Re-design project of the Istat consumer price survey: use of probability samples of scanner data for the calculus of price indices

Antonella Bernardini, Maria Cristina Casciano, Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Giovanni Seri, Marco Dionisio Terribili, Francesca Tiero [1]

Abstract

The availability of scanner data (SD) from the retail modern distribution (food and grocery) is the starting point for the implementation of the innovation in the consumer price survey (CPS) improving and unburdening the data collection phase, together with the progressive introduction of more rigorous probabilistic sampling procedures for the selection of outlets and products (series). This article presents the work carried out by the statistical-methodological working group for the revision of the sample design of the consumer price survey in the light of the new data sources (SD). The experiments of probabilistic selection schemes of the series are developed in two main phases that reflect the choice made by Istat to make a gradual transition from a fixed approach to a dynamic approach in the calculation of the consumer price index (CPI). In 2018, Istat started using the SD to compute the CPI.

Keywords: scanner data, selection schemes, price indices, fixed and dynamic approach.

1 Antonella Bernardini (anbernar@istat.it); Maria Cristina Casciano (casciano@istat.it); Claudia De Vitiis (devitiis@istat.it); Alessio Guandalini (alessio.guandalini@istat.it); Francesca Inglese (fringles@istat.it); Giovanni Seri (seri@istat.it); Marco Dioniso Terribili (terribili@istat.it); Francesca Tiero (tiero@istat.it), Italian National Institute of Statistics – Istat.

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics – Istat.

1. Introduction

The Consumer Price Survey (CPS) has undergone for some years a review aiming at redesigning different aspects of the data collection procedures and sampling methods. The aim of the project is to introduce progressively in the CPS more rigorous sampling procedures, probabilistic where possible, starting from the selection of outlet and products for the sectors where this is feasible.

The availability of scanner data (SD) from the retail modern distribution (food and grocery) is the starting point for the implementation of the innovation in the Consumer Price Survey. At present food and grocery sector cover 11% of the total and modern distribution the 55% of the total. Scanner data is a big opportunity to introduce improvements in the CPS both for the data collection and the sampling perspective.

In recent years Istat, through a contract with Nielsen and an agreement with the six main retail chains operating in Italy, started, at the end of the year 2014, receiving SD referred to food and grocery market and processing them to experiment the calculus of Consumer Price Index (CPI). This acquisition places Italy among countries using or testing the use of this source of data for compiling CPI. As noted some years ago in the ILO CPI Manual (ILO *et al.*, 2004), "Scanner data constitute a rapidly expanding source of data with considerable potential for CPI purposes" (p. 54); "Scanner data obtained from electronic points of sale include quantities sold and the corresponding value aggregates on a very detailed level" (p. 92); "Scanner data are up to date and comprehensive" (p. 478).

After an experimental phase, since 2018 Istat has started using the SD concerning grocery products (processed food, products for house maintenance and personal care) to compute the consumer price index (CPI). SD are regularly provided to Istat through the market research company ACNielsen, for the main chains present in Italy and a sample of about 2,000 outlets deployed all over the country. For 2018 the compilation of the CPI using scanner data is based on a fixed basket perspective: the use of these data has concerned some channels of the modern distribution, in particular hypermarkets and supermarkets of the main retail chains operating in Italy. Since 2020 Istat has extended the use of scanner data to other channels of the modern distribution (discounts, small sales areas and specialist drug) and has realised the transition to a flexible basket approach. At the moment, the traditional retail distribution data continue to be collected by the current survey on the field. This choice is determined by the peculiarities of the Italian retail distribution with respect to other countries: the traditional distribution, in fact, is still relevant in many geographical areas of the country.

This paper aims to present the work carried out by the statistical-methodological group for the revision of the sampling design of the consumer price survey in the light of the new data sources (SD). The work is divided into two main lines of research, the first concerning the analyses of the scanner data acquired from 2014 by Istat for the experimental phase, the second involving the study of selection schemes of series (references individuated by EAN and outlet codes) from SD. In the paper, the focus is both on the results of the analysis conducted on the big data sets containing SD and on the experiments in the use of samples from scanner data for the calculus of elementary indices for homogeneous product aggregates (consumption segments, markets).

The analyses conducted on scanner data range from the aggregate level, *i.e.* statistical distributions of turnover by chain and province, to the elementary level, *i.e.* continuity and seasonality of the products and data quality. These analyses have been possible thanks to the information contained in SD. Scanner data files, indeed, contain elementary information referred to single EAN codes (GTIN - Global Trade Item Number, or EAN - European Article Number) for specific outlets consisting of turnover and quantities sold during a week. This information does not provide the "shelf price" of the product individuated by the EAN code and outlet (references or series) but allows us to define a unit value or average weekly price.

For SD experimental framework, firstly, probability and nonprobability selection schemes of series are compared and then different probability sampling designs are examined. The experiments are conducted to test different methods for selecting references by comparing the estimates of different elementary index formulas with the value of the corresponding price indices obtained from the whole set of references for the same elementary aggregate. Sampling designs and price index formulas are studied through a Monte Carlo simulation by first selecting 500 samples for each different sampling design and then calculating variability and bias on the estimated indices in the replicated samples.

Moreover, through a further experiment, the differences between a fixed and a dynamic population approach in the construction of the elementary price indices are highlighted. In this case, the purpose is trying to measure the magnitude of sampling and non-sampling errors for different price index formulas in both approaches. These errors are generated by different causes, among the main ones: disappearing products, ignoring entries of new products and temporary missing products, when a static population is assumed and a fixed basket is used; variability in the number of matched-items between the months and the chain drift of weighted price indices when a dynamic population is considered and a flexible basket is used.

In both cases the experiments were conducted starting from a sampling frame represented by a panel data set that contains permanent series (references), which refers to those series with not-null turnover for at least one relevant week (the first three full weeks) in each month of the considered year, starting from the December of the previous year. The use of data referred to the full weeks (first two, or first three) for each month is advised in the recommendation drafted by Eurostat. This is a restriction inherent to the observable weeks that arises from reasons deriving from operational constraints of the productive process.

In the paper, the use of scanner data in the consumer price index production (Section 2) and the Istat project of redesigning the CPI are presented (Section 3). Then, in Sections 4 and 5, the processing phases and the analyses conducted on scanner data are shown, while in Section 6 the experimental framework for sampling from SD is described. Sections 7 and 8 present a detailed description of the experiments and the main results. Finally, Section 9 describes the sampling design adopted by Istat from 2018 to 2020 to select the sample of outlets from the SD of grocery products and how it came to its definition. Conclusions are reported in Section 10.

2. Scanner data in the Consumer Price Index production

2.1 Practice in the European countries

In a study perspective scanner data from retail stores allows researchers to evaluate how different price index formulas perform at the elementary level. In fact, official CPI is usually constructed in two broad steps. First, elementary price indices are calculated for narrowly defined and relatively homogeneous products, known as elementary aggregates. In a second step, these elementary indices are aggregated into a single consumer price index using expenditure weights. Elementary indices, named also higher-level elementary indices, are therefore the building blocks of price index numbers.

While the aggregation at a higher level is carried out using generally Laspeyres type formulas with weights deriving from national account or expenditure survey data, official practices in elementary price index construction are still not uniform across countries, deserving further investigation in the consequences of different choices (Gábor and Vermeulen, 2014).

The launch of barcode scanner technology has enabled retailers to capture detailed information on transactions at the point of sale. Scanner data is high in volume and contains information about individual transactions or summaries, date, quantities and values of products sold, and product descriptions. As such it is a rich data source for NSIs that can be used both to improve their statistics and to reduce statistical burden and costs.

Scanner data will be increasingly available to statistical agencies and consequently, new methods are needed to work with this new data source. In Europe, countries are rapidly expanding the use of SD for the compilation of the Consumer Price Index (CPI). Norway was the first country using scanner data in regular CPI production (2001) followed by the Netherlands. Other countries have also started using scanner data, for instance, Sweden, Switzerland, Belgium, and Denmark. Scanner data was introduced in the regular production of the CPI from January 2018 in Luxemburg and Italy. From 2016, other NSIs (as CSO of Ireland) have started projects to obtain and analyse scanner data from retailers to research its potential use in the production of statistics and the CPI in particular.

Scanner data can be exploited in different ways. The simplest way is using SD as an alternative source for price collection, replacing collection within the stores, without changing the traditional principles of computing the price indices. This method was applied by the Swiss Federal Statistical Office (Vermeulen and Herren, 2006). Alternatively, as done in Norway and Sweden, SD can be used as the universe from which samples of references can be selected following different methods (Nygaard, 2010; Norberg, 2014).

Finally, all (or almost all) SD can be used to compile price indices, without a strict sample selection, but with consequences on the theoretical definition of the index. In Belgium and Netherlands, the computation method is different and the data are used in a more extensive way to calculate price indices (van der Grient and de Haan, 2010). The method used assumes a dynamic population approach: elementary price indices of homogeneous items are calculated by monthly chained unweighted geometric index (Jevons); no explicit weighting is applied and expenditure information is used just to select a cut-off sample of matched items during two months in a row.

2.2 Impact on Consumer Price Index compilation

Scanner data introduce important advantages compared to data collected through traditional survey. In particular, the availability of turnover and quantity data at the item level offers a real possibility of calculating more accurate indices: it is possible, in fact, to include in the calculus the expenditure share of each product sold. SD also contains descriptive information about items characteristics useful to treat quality change, to identify relaunches of existing products or new products, etc. (Feldmann, 2015; Chessa *et al.*, 2017).

On the other hand, the use of SD in the compilation of CPI must take into account some important drawbacks, as attrition of products, temporary missing products, entry of new products and volatility of the prices and quantities due mainly to sales. These are aspects that need to be addressed from both a theoretical and a practical point of view (de Haan *et al.*, 2016).

To maximise the potential offered by SD it would be necessary to go beyond those methods of price index compilation which do not exploit all the information provided by the data and do not take into account the population dynamics (Chessa *et al.*, 2017). Weighted and chained indices should be considered to incorporate the overall price trend over a given time, including the prices of new products. Furthermore, the problem of shrinkage over time due to the attrition of a fixed basket of products is solved automatically using chain indices. However, even though in a dynamic approach it is necessary to construct series of chained indices, high-frequency chaining of weighted indices (also superlative Fisher and Törnqvist indices) are affected by chain drift, due to non-symmetric effects on quantities sold and expenditure share of goods before and after the sale (Ivancic *et al.*, 2011; de Haan and van der Grient, 2011).

In recent years, an important debate has taken place among the researchers dealing with the estimate of the consumer price index starting from SD. The focus, above all, has been on the transition from a static population approach (fixed basket) to a dynamic population approach (flexible basket) and it is based on the study of alternative price index formulas based on matched-model methods (matching of products sold during two months in a row) or other methods that are transitive and, therefore, free from chain drift (de Haan *et al.*, 2016).

Other aspects discussed are the quality of SD (completeness and correctness) and the definition of methods to treat appearing and disappearing products, temporary missing products, relaunches, quality change, etc. (Vermeulen and Herren, 2006; van der Grient and de Haan, 2010).

3. Re-Design of the Italian Consumer Price Survey

The aims of the re-design of the Italian Consumer Price Survey are to be ascribed to a reduction of the weight of the traditional data collection in the field to around 50% through expanding the use of the Internet as a data source, widening the use of web scraping techniques and using scanner data as a new source. Scanner data are a great opportunity for introducing probability sampling designs: the selection of elementary items from scanner data allows to implement a sufficiently feasible field procedure and to overcome the potential source of bias of the procedure adopted in the current survey, based on subjective choices[2].

To introduce the use of the SD in the compilation of the CPI, Istat's contract with Nielsen initially provided for the supply of weekly data of turnover and quantities at EAN code (elementary item) and outlet level for six modern retail distribution chains operating in the food and grocery market in 37 Italian provinces (coverage of 55% of the Italian population). For the experimental phase, Nielsen provides backward data for at least one full year and the preceding month of December, starting from December 2013, 2014 or 2015 (depending on the starting point of delivery of each province).

During 2014 and 2015 the scanner data were provided by Nielsen gradually, first the data relating to five provinces, to which were added then 14 and 18 other provinces up to a total of 37 provinces. The release of scanner data by Nielsen follows an informal agreement between Istat and the Association of Modern Distribution, representing the main chains of modern retail trade. The requests of data (with formal and legal meaning) sent by Istat concerned weekly data of turnover and quantities, EAN code, outlet code and others information, for a progressively increasing amount of provinces, for the six

2 The traditional Consumer Price Survey (CPS) carried out at territorial level is based on three purposive sampling stages. The sampling units are respectively the municipalities, the outlets and the elementary items for which the prices are collected. The biggest municipalities are forced by law to participate to the survey. The Municipal Offices of Statistics select the outlets sample, where the prices of a fixed basket of products (including roughly 1,000 products) are collected. The outlets sample is chosen to be representative of the consumer behaviour in the municipality. For each product of the basket the most sold item is selected and the prices of these items are collected throughout the year. At the end of each year Istat refreshes the fixed basket of products and, at the same time, the sample of outlets and elementary items is updated. The elementary price indices are currently obtained at municipality level by unweighted geometric mean. The general price index is calculated by subsequent aggregation of elementary indices, using weights at different levels based on population and national account data on consumer expenditure.

"big chains" (Coop Italia, Conad, Selex, Esselunga, Auchan, Carrefour) covering almost 57% of the total turnover of modern distribution.

Moreover, Nielsen provided the dictionary for the classification of EAN code sold in Italy attributes that allow you to identify the product (manufacturer, brand, possible sub-brand, size, packaging, variety) to GS1-ECR-Indicod product classification (variation of GPC Global Product Classification applies worldwide). Istat ensures internally the translation from ECR to COICOP, the classification of products used for the CPI. Consumption segments, not foreseen by the EU-COICOP, are the most detailed domain of estimate for Italian CPI, constitute groupings of homogeneous products; those defined for the food and grocery are 126 out of a total of 324.

4. Processing phases of scanner data

4.1 Outline

The analyses on scanner data quality constituted an important activity of the statistical-methodological workgroup aimed at guaranteeing the completeness and correctness of the acquired data. Completeness and correctness of the data are two important pillars for the correct use of scanner data (SD) and the compilation of an accurate modelling of the price index over time.

These analyses are part of the processing phase of the collected data in which the quality checks must be performed and the data cleaning must be made.

Indeed, the use of SD implies the definition and the implementation of different checks in both data acquisition and processing phases. The scanner data collection requires the use of formal checks on the flow of collected data but also the development of checks on the quality of the data. The formal checks must be defined to ensure the completeness of the data collected at the provincial level, distribution chains, outlets, products and weeks. The quality checks on loaded data are required to introduce editing rules which identify inadmissible values on the variables of interest as quantities sold, turnover and prices (Rais, 2008; Saidi and Rubin Bleuer, 2010).

As the scanner data file do not provide the “shelf price” of the product individuated by the EAN code and outlet (references or series), the unit prices are defined as a unit value or average weekly price of series starting from the quantities sold and the turnover.

4.2 Formal and quality checks

The continuous flow of acquired scanner data by Nielsen is subject to formal checks both during and after data loading.

The formal checks during the loading of data affecting the presence of a full numeric code for the outlets and products, and the presence of numeric and valid decimal values for the turnover and quantities. Another important check affects the week in which the data must refer to.

The formal checks of the post-loading concern:

i) the presence of duplicates for outlet, reference and week;

ii) the absence of an outlet or a product in the list updated every six months in the first case and every two months in the second case;

iii) the presence of null fields of turnover and quantities;

iv) the presence of unauthorised data such as outlets that do not belong to the authorised provinces or the allowed chains, or outlets not classified as a supermarket or hypermarket.

These checks must ensure that data always refer to the same population (provinces, chains, outlets and products) for each week.

The data quality check follows the phase of formal checks with the aim of introducing editing rules that identify inadmissible values among the variables of interest (quantities sold, turnover and unit prices). First quality checks are implemented to identify and eliminate the problematic occurrences (outlet, EAN code, week) in which:

v) quantity<1 not motivated by unit of measurement;

vi) decimal values on quantities>1 not motivated by unit of measurement;

vii) unit prices <= 0.01 €.

Subsequently, in order to maintain an accurate price index over time, the product prices have been validated considering the data acquired, during the relevant weeks of every month, at the provincial level.

In the following schemes, the formal checks developed during and after the upload are synthesised.

Scheme 1 - Formal checks during the upload

Check	Error type
ID outlet	Not null and numeric
ID product	Not null and numeric
Turnover	Numeric
Quantity (sold packages)	Numeric
Year	Current or previous year
ID week	Not ≥ 53

Scheme 2 - Formal checks after the upload

Check	Error type
Duplication	1 - Duplicated outlets, products or weeks
Outlet	2 - Outlet not in the frame of outlets (updated every 6 months)
Product	3 - Product not in the frame of products (updated every 2 months)
Missing turnover/quantity	4 - Turnover and/or quantity NULL
Quantity	51 - quantity <1 52 - values with decimals
Prices	61 - prices ≤0.01 €
Completeness	71 - Data of the authorised weeks 72 - Data from authorised provinces 73 - Data from authorised chains 74 - Type of outlets (only Hypermarket and Supermarket)

4.3 Identification of inadmissible unit prices

To identify inadmissible unit prices, the price of an occurrence (product*outlet*week) too high or too low concerning an interval built on a synthetic measure of the distribution of the prices of the product sold in the province in the month, several methods have been tested. For the sake of simplicity, also in terms of computational burden, the choice has been reduced between the two methods. Both methods are based on the computation of the median unit prices, considering the quantities sold for each single occurrence (EAN code, week, outlet).

The first method consists of a fixed trimming method and the tolerance interval of prices is:

$$\left(\frac{Median_w}{K_1}, K_2 * Median_w\right); \qquad (1)$$

the second can be named moving trimming and the related tolerance interval is:

$$\left(Median_w - \frac{K_1 - 1}{K_1}\frac{Median_w}{log_{10}\left(Median_w + 10\right)}, Median_w + \left(K_2 - 1\right)\frac{Median_w}{log_{10}\left(Median_w + 10\right)}\right). \qquad (2)$$

Trimming depends on the values assigned to K_1 and K_2. Assigning values to K_1 and K_2 requires making assumptions on maximum discount and maximum price hike with respect to the median price of the product allowed. For food and groceries, $K_1=5$ (it means that up to 80% discount is allowed) and $K_2=3$ (it means that up to 3 fold increase with respect to the median price of the product is allowed) seems to be plausible values.

However, while in fixed trimming the relative width of the tolerance interval remains unchanged, with moving trimming it narrows as the median price of the product increases (Figure 1 and Table 1).

Figure 1 - Tolerance interval limits of (prices/Median$_w$) with fixed and moving trimming

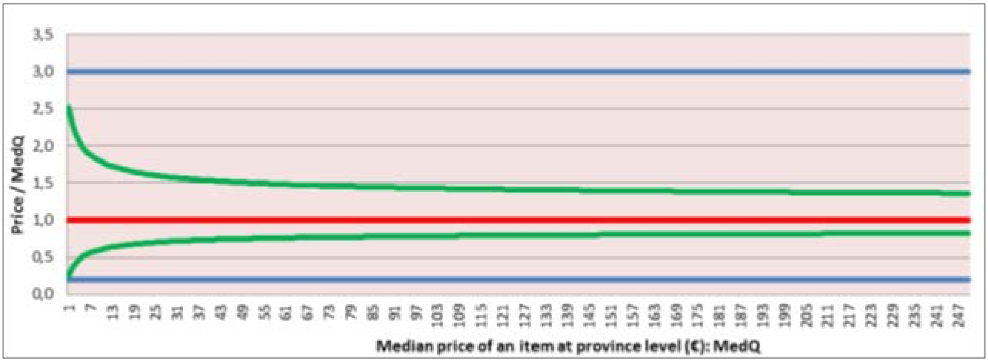


Table 1 - Lower (LB) and Upper (UP) bounds calculated with two trimming methods on a Median price

Median price	Fixed trimming		Moving trimming	
	LB	UP	LB	UP
1	0.2	3	0.246	2.523
2	0.4	6	0.362	2.289
3	0.6	9	0.432	2.147
4	0.8	12	0.48	2.05
5	1	15	0.516	1.979
6	1.2	18	0.543	1.924
7	1.4	21	0.565	1.88
8	1.6	24	0.583	1.843
9	1.8	27	0.598	1.813
10	2	30	0.611	1.786
20	4	60	0.683	1.64
25	5	75	0.702	1.602
50	10	150	0.75	1.504
75	15	225	0.773	1.459
90	18	270	0.781	1.442
100	20	300	0.786	1.432
120	24	360	0.794	1.416
200	40	600	0.814	1.377

The moving trimming is preferable to the fixed one because it narrows down, thanks to the log function, the extremes as the median price of the product increases.

4.4 Removal of inadmissible unit prices effect

In this paragraph, some results obtained from the removal of the occurrences identified with the moving trimming method are presented.

The analysis reported here was conducted on all occurrences by chains available (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) in the years 2013 and 2014, in five Italian provinces. The analysis aims to highlight the impact that the removal of inadmissible unit prices has in terms of the number of occurrences, the number of quantities sold and turnover, and to detect the consumption segments most affected by the deletion of occurrences.

Table 2 presents the removal of inadmissible unit prices effect on the total amount of occurrences, quantities sold and turnover.

Table 2 - Total amount of occurrences, quantities sold and turnover – percentage of references, quantities sold and turnover removed, by year and province

Province	Year	Total amount of			Removal of inadmissible unit prices effect in terms of		
		Occurences	Quantities sold	Turnover	Occurences (%)	Quantities sold (%)	Turnover (%)
Ancona	2013	16,314,683	179,261,691	302,928,869	0.042	0.027	0.024
	2014	16,972,117	183,593,696	309,328,337	0.033	0.021	0.019
Cagliari	2013	11,236,941	145,630,320	256,859,395	0.044	0.027	0.028
	2014	11,383,752	145,612,723	253,942,082	0.044	0.032	0.030
Palermo	2013	10,671,436	139,756,595	222,335,145	0.064	0.030	0.026
	2014	12,094,184	152,094,573	240,512,170	0.098	0.033	0.030
Piacenza	2013	6,474,872	93,362,079	167,777,673	0.021	0.009	0.017
	2014	7,521,222	100,610,744	180,572,592	0.030	0.020	0.027
Torino	2013	45,458,148	671,423,702	1,242,080,444	0.048	0.017	0.023
	2014	48,621,536	679,459,389	1,248,849,275	0.062	0.026	0.026

Generally, the number of removed occurrences is very low (Table 2) in all provinces but not constant in the two years under review, except the province of Cagliari in which the deleted occurrences is 0.044 percent in both years.

The turnover share lost due to the elimination of occurrences is very contained, in fact, is never more than 0.030 percent.

By analysing the problem for a single consumption segment, the situation changes as some consumption segments lost 0.05 percent of turnover.

In scheme 3 are listed, for each province and each year, the consumption segments that suffer a greater loss of turnover share following the removal of inadmissible unit prices related to specific occurrences.

Scheme 3 - Consumption segments (COICOP-6digit) with percentage of turnover removed >0.05%

ANCONA	2013	Mineral water; Other cereal-based products; Non-alcoholic beer, or beer with low alcoholic content; Cured cheese; Dried fruit; Berries; Ice cream; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Rice
	2014	Mineral water; Other meat; Other cereal-based products; Non-alcoholic beer, or beer with low alcoholic content; Body, hand and hair lotions; Ready meals with ground meat; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products
CAGLIARI	2013	Mineral water; Other beauty products; Other medical products; Other products for pets; Body, hand and hair lotions; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Salt, spices and aromatic herbs; Sugar
	2014	Mineral water; Frozen seafood; Body, hand and hair lotions; Dried, smoked or salted fish or seafood; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Pregnancy tests and contraceptives; Sugar
PALERMO	2013	Other alcoholic beverages; Other disposable items for the home; Other perishable items for the home.; Other preserved fish or seafood; Other beauty products; Other products for pets; Detergents and house cleaning products; Nuts; Dried fruit; Ice cream; Body, hand and hair lotions; Fresh pastry; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Rice; Brushes, brooms, wipes and sponges; Dried vegetables; Sugar
	2014	Mineral water; Other beauty products; Other house cleaning and upkeep products; Non-electric appliances; Soaps and personal hygiene products; Detergents and house cleaning products; Dried fruit; Body, hand and hair lotions; Packaged bread; Frozen fish; Pizza and quiche; Fresh pastry; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Rice; Brushes, brooms, wipes and sponges; Gourmet wines; Sugar
PIACENZA	2013	Fresh Milk
	2014	Mineral water; Non-alcoholic beer, or beer with low alcoholic content; Fresh milk; Paper-based kitchen products
TORINO	2013	Natural meat and mixed-meat hamburgers; Perfumes and make up products; Electric shavers and trimmers; Mineral water; Dried, smoked or salted fish or seafood; Other meat; Dried vegetables; Body, hand and hair lotions; Other house cleaning and upkeep products; Non-electric appliances; Pregnancy tests and contraceptives
	2014	Mineral water; Other products for pets; Other house cleaning and upkeep products; Non-electric appliances; Body, hand and hair lotions; Dried, smoked or salted fish or seafood; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Brushes, brooms, wipes and sponges; Alcoholic beverages; Pregnancy tests and contraceptives

From the scheme above it is clear that the consumption segments like *mineral water, perfumes* and *razors* are more problematic than other ones, in fact, in this segment, a greater number of references with inadmissible unit prices can be identified.

4.4 Moving trimming method

In the moving trimming method, as said above, the relative allowed range shrinks contextually with the increase of the median price of the product (Figure 1). In the graphs below (Figures 2-6), in which the density functions of the unit prices of some specific products are represented, the tolerance intervals are delimited by two vertical red lines, while the green line and the dotted green line indicate respectively the weighted and unweighted median price of the product. The values of inadmissible unit prices are indicated by red dots; the green dots indicate the presence of discounts.

Figure 2 - Density function of the unit prices of product "Powdered milk for babies Danone"

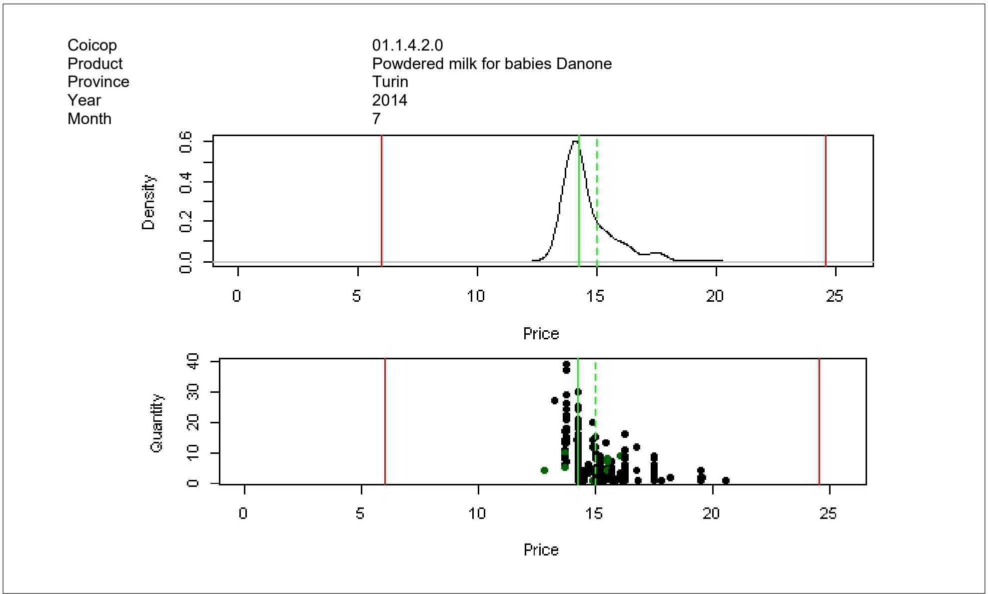


Figure 3 - Density function of the unit prices of product "Aged rum, Havana Club, 700 ml"

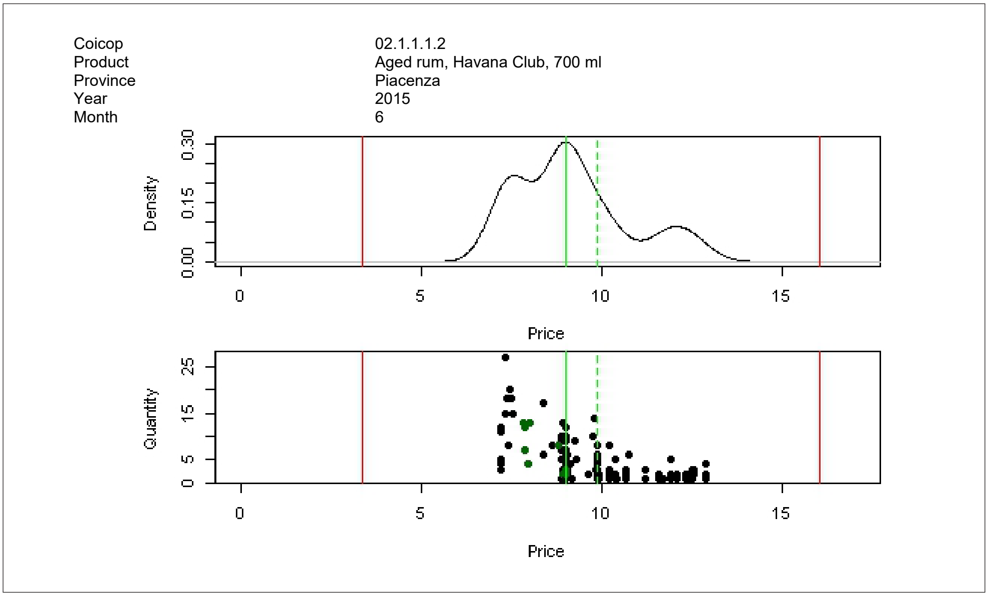


Figure 4 - Density function of the unit prices of product "Rice, Private Label, 1000 gr"

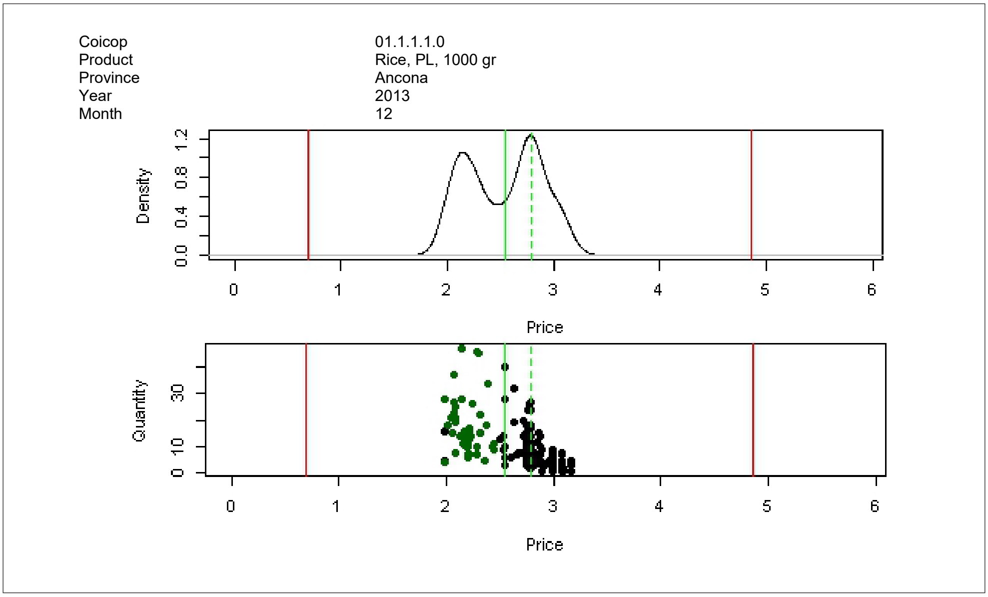


Figure 5 - Density function of the unit prices of product "Sparkling water, Rocchetta, 1.5 lt"

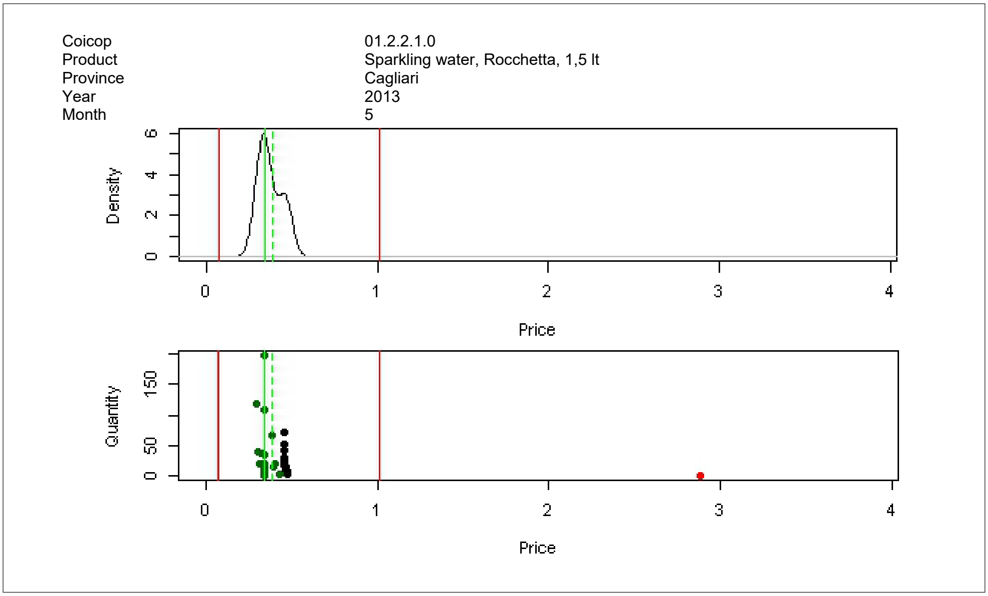
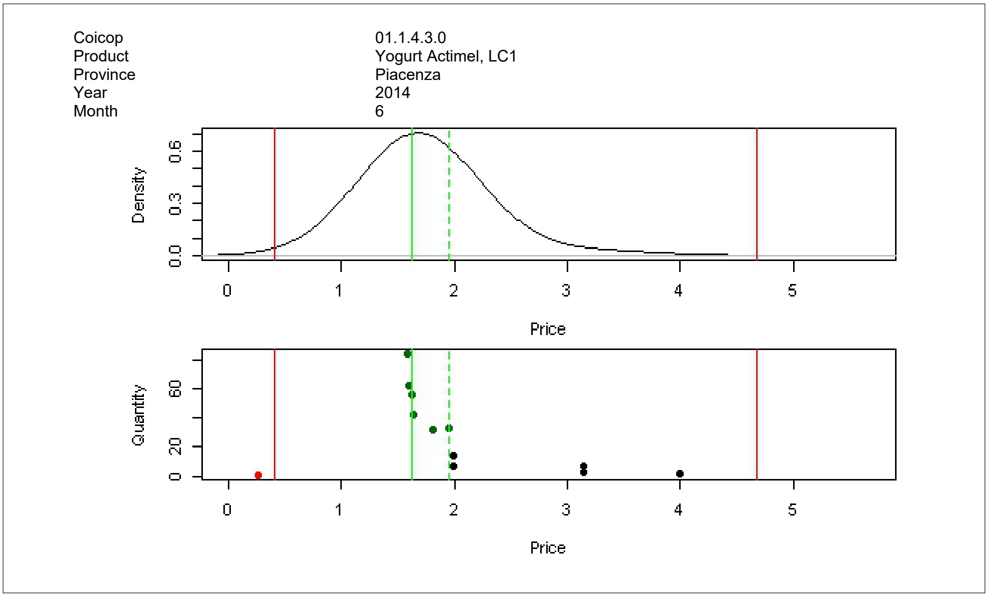


Figure 6 - Density function of the unit prices of product "Yogurt Actimel, LC1"



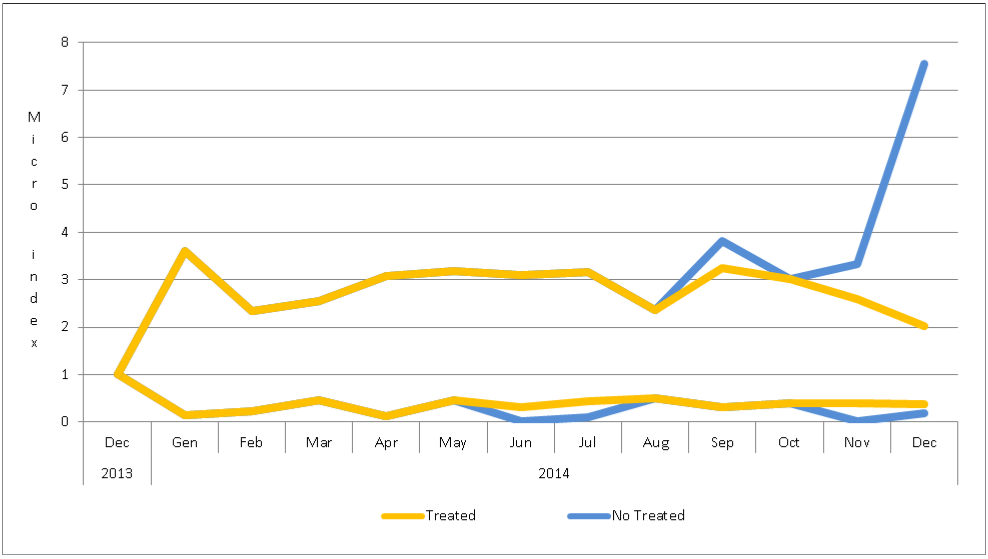
The method does not reveal any inadmissible value in unit prices in graphs 2, 3 and 4, while in graphs 5 and 6 identify the inadmissible values respectively beyond the upper and lower limit of tolerance interval.

4.5 Impact of data cleaning on micro-indices

For an assessment operated by the data cleaning, an analysis on micro-indices - the ratio between the price of the occurrence at time t and the price of the same occurrence at the base time - has been performed. In particular, the micro-indices trend has been observed in the twelve months of 2014.

Figure 7 compares the maximum and minimum value of micro-indices calculated before and after the phase of the data processing in the consumption segment wine of the Torino province.

Figure 7 - Micro-indices trend before and after the phase of the data processing (2014)



The figure above shows that in some months the impact of data cleaning can be quite important, besides its conceptual correctness: in the examined case the difference between the maximum values of the micro-indices of treated and untreated data, in December, is remarkable.

5. Analyses on scanner data

5.1 Framework

The analysis of scanner data has constituted an important line of research conducted by the statistical-methodological workgroup especially in the initial phase of the project. The realised analyses had different objectives:

a) to study the chain and outlet type (hypermarkets and supermarkets) distributions in the first provinces acquired in terms of turnover;

b) to evaluate the attrition problem related to the life cycle of the EAN code and series (or references);

c) to study the continuity of the presence of the EAN codes and temporary availability (such as seasonal, new entry, temporary or definitive absence);

d) to identify seasonal products (products sold just at certain times of the year, following a seasonal pattern) and to study seasonality in consumer prices caused by a variety of influences, some on the demand side and some on the supply side.

The analyses conducted on the big data sets acquired by Istat and analysed herein, cover five Italian provinces (Ancona, Cagliari, Palermo, Piacenza, Torino) and six chains of modern distribution (Conad, Coop, Esselunga, Auchan, Carrefour, Selex). The analyses are focussed on the observed series (EAN+outlet code) belonging to the relevant weeks (the first three full weeks in each month) of each month and on permanent series (panel series SD), as defined above.

The analyses on continuity and seasonality of the products were conducted only in Torino province.

5.2 Distribution of turnover and outlet

In the following analysis, some aspects of the six chains of modern distribution (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) and outlet type (hypermarket and supermarket) distributions in the first five provinces acquired by Istat are highlighted for the year 2014.

Table 3 - Total turnover and number of outlets by chain and province (2014)

Chain		Province					
		Ancona	Cagliari	Palermo	Piacenza	Torino	Total
Conad	Turnover	14,287,546	63,553,952	56,233,579	38,883,578	47,924,071	220,882,726
	Outlet	3	10	15	7	13	48
Coop	Turnover	79,792,735	-	35,976,469	32,823,101	272,443,053	421,035,357
	Outlet	10	-	2	4	26	42
Esselunga	Turnover	-	-	-	58,935,833	108,143,835	167,079,668
	Outlet	-	-	-	2	3	5
Auchan	Turnover	106,363,959	54,247,839	64,424,104	4,890,904	157,247,092	387,173,897
	Outlet	11	2	7	1	8	29
Carrefour	Turnover	11,093,852	38,062,177	64,896,292	2,659,154	424,278,376	540,989,851
	Outlet	1	2	21	1	40	65
Selex	Turnover	74,711,322	77,908,712	-	28,197,290	140,878,602	321,695,925
	Outlet	34	21	-	4	39	98
Total	Turnover	286,249,412	233,772,680	222,097,542	166,389,859	1,152,108,585	2,058,857,425
	Outlet	59	35	46	19	130	289
Coverage turnover		87.00	74.28	69.95	73.68	72.65	73.96

The analysis has been carried on the whole of the 289 outlets of the provinces during the 52 weeks of the year 2014. Table 3 contains the whole turnover and the number of outlets by chain and province and the percent coverage of the six chains with respect to the total turnover of modern distribution for food and grocery at the province level.

The table above shows a heterogeneous situation both among the chains and the provinces: Torino province represents more than 50% of turnover involved; this remark could be influenced by the high number of outlets observed in this province, 130 on a whole set of 289. In the other provinces the number of outlets varies from 19 (Piacenza) to 59 (Ancona), with a turnover between 166 and 286 million euro.

The last row shows the good level of coverage of the six chains, although with a certain heterogeneity among provinces: the coverage is generally close to 72%, with a maximum level of 87% assessed in Ancona province.

In Table 4, likewise above, turnover and number of outlets are reported considering the six chains and two outlet types. This variable is particularly important because connected both with chain (which can have a higher/lower propensity to set up a hypermarket/supermarket) and with the number of sold elementary items, much higher in the hypermarkets than in the supermarkets.

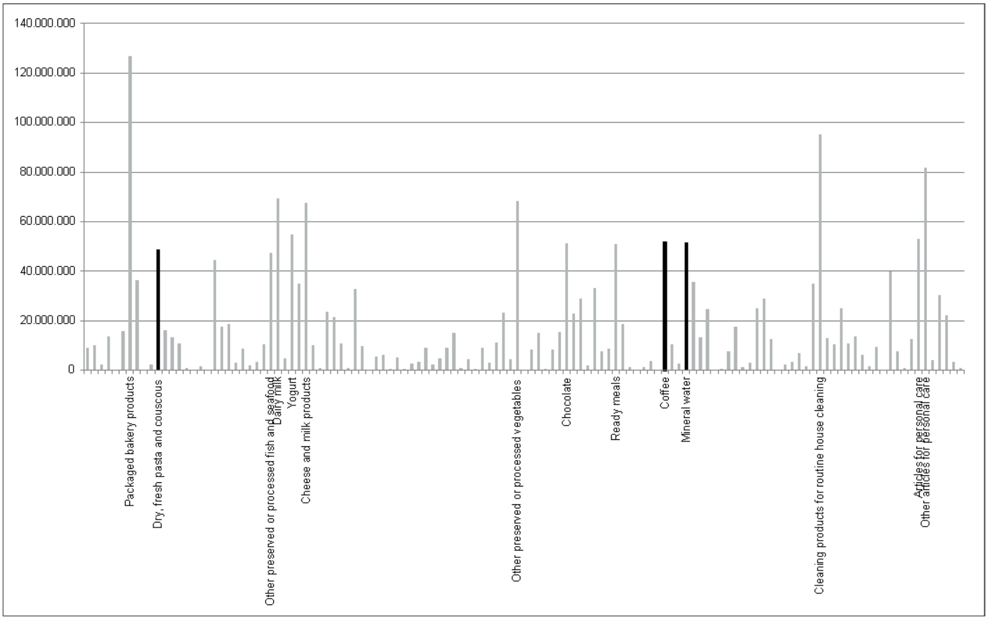
Table 4 shows a greater hypermarkets turnover than supermarkets one, although the number of outlets is reasonably lower (53 hypermarkets and 234 supermarkets, for a total of 289 outlets). Moreover, hypermarkets are more spread in the Torino province (28 outlets) with respect to Palermo and Piacenza provinces (4 outlets). This result, however, has to be linked to the distribution of the local chain, in fact, Esselunga and Carrefour chains have a greater number of hypermarkets than the other chains.

Table 4 - Total turnover, number of outlets and number of items by province and outlet type (2014)

Province		Outlet type		Total
		Hypermarket	Supermarket	
Ancona	Turnover	136,000,290	150,249,122	286,249,412
	Outlet	10	49	59
	Item	63,112	43,000	106,112
Cagliari	Turnover	126,001,578	107,771,102	233,772,680
	Outlet	7	28	35
	Item	53,053	26,418	79,471
Palermo	Turnover	81,250,169	140,129,020	222,097,542
	Outlet	4	41	46
	Item	40,532	33,349	73,881
Piacenza	Turnover	83,803,482	82,586,377	166,389,859
	Outlet	4	15	19
	Item	44,341	50,466	94,807
Torino	Turnover	738,418,729	412,833,155	1,152,108,585
	Outlet	28	101	130
	Item	83,342	54,420	137,762

The following Figure 8 shows the total turnover of all outlets of the five provinces for each of the 126 consumption segments (6-digit COICOP classification) for food and grocery. The wide range of the consumption segment turnover arises from the figure. The black bars highlight the three consumption segments (coffee, pasta and mineral water) on which the first experiments of the selection of the samples were concentrated even if only for the Torino province.

Figure 8 - Total turnover (six chains and five provinces) for consumption segment (2014)



5.3 Distribution of turnover - relevant weeks and permanent series

In this paragraph, the analyses are focussed both on the observed series (EAN+outlet code) belonging to the relevant weeks of each month and on permanent series (panel series SD).

Table 5 shows the whole turnover observed for the five provinces, respectively on the whole set of series (A), on the relevant week series (B) and the panel series (C).

Table 5 - Total turnover for all series, relevant week series and panel series, five Italian provinces (2014)

Province	Turnover			% Coverage		Number of panel series
	All weeks all series (A)	Relevant weeks all series (B)	Relevant weeks panel series (C)	B/A	C/B	
Torino	1,152,108,585	793,433,903	562,758,215	68.87	70.93	7,185,048
Ancona	286,249,412	199,336,585	134,533,396	69.94	67.49	2,331,144
Cagliari	233,772,680	162,008,273	109,160,452	69.30	67.38	1,583,448
Palermo	222,097,542	153,924,950	91,512,573	69.31	59.45	1,388,724
Piacenza	166,389,859	115,388,514	82,736,445	69.35	71.70	1,123,092

Observing only relevant weeks allows one to take into account about 70% of the turnover of all weeks (52 weeks of the year 2014), without relevant local differences. Then, looking at the coverage turnover of the panel series with respect to relevant week series, it ranges from 59.45% (Palermo) to 71.70% (Piacenza).

5.4 Continuity of products

Following the underlying idea that price indices will be computed on the base of a fixed sample (basket) of series during a twelve months period, in this paragraph, the continuity of the presence of the EAN codes in general (elementary items regardless of the outlets where they are sold) in the period is investigated in order to help in determining relevant subsets of items to be included in the basket and in the population to be sampled.

Prices of the series (EAN+outlet code) included in the basket are collected each month and therefore they should be all available at each time unless products acknowledged as seasonal. The cycle of life of the elementary items (EAN codes) is then investigated in order to identify relevant seasonal products other than those already known as seasonal (Fruits and Vegetables for example).

A definition of continuous EAN code is then needed and, on the complementary side of the non-continuous EAN code, it has to be distinguished between different kinds of temporary availability (such as seasonal, new entry, temporary or definitive absence).

The availability of the item/product for price collection can be registered for each of the 52 weeks in a year. Moreover, for each month the first 3 complete weeks are regarded as relevant for price collection.

Four different definitions of continuous items are considered:

a) available at least in 1 relevant week for each month;

b) available at least in 2 relevant weeks for each month;

c) available in all the 3 relevant weeks for each month;

d) available each of the 52 weeks in the year.

It worth notice that the four different definitions define a hierarchy in the sense that the set of continuous products according to the d) is included in the one obtained applying each of the previous definitions, and so on.

In Table 6 the number of different items registered in Torino during 2013 is reported for each of the above definitions. Moreover, the coverage with respect to the total number of items and to the amount of turnover is shown in percentage.

In all four cases the set of continuous EAN codes represents more than 80% of the total amount of the turnover. Going through the definitions from a) to d) the number of different items decreases more than 10% while the coverage of the total amount of the turnover decreases of less than 4%.

The definition a) has been then adopted to identify the set of continuous products whose relevance and characteristics are investigated in the next paragraph. The attention here is focussed on the complementary set of non-continuous EAN codes.

Table 6 - Number of continuous items, coverage of the total number of items and of the total amount of turnover by four different definitions (Torino, 2013)

Availability	N. Item	% Coverage	
		Total item	Total turnover
At least one week each month	47,760	49.7%	87.0%
At least two weeks each month	43,499	45.3%	85.5%
Three weeks each month	38,469	40.1%	83.6%
52 weeks a year	37,450	39.0%	83.1%

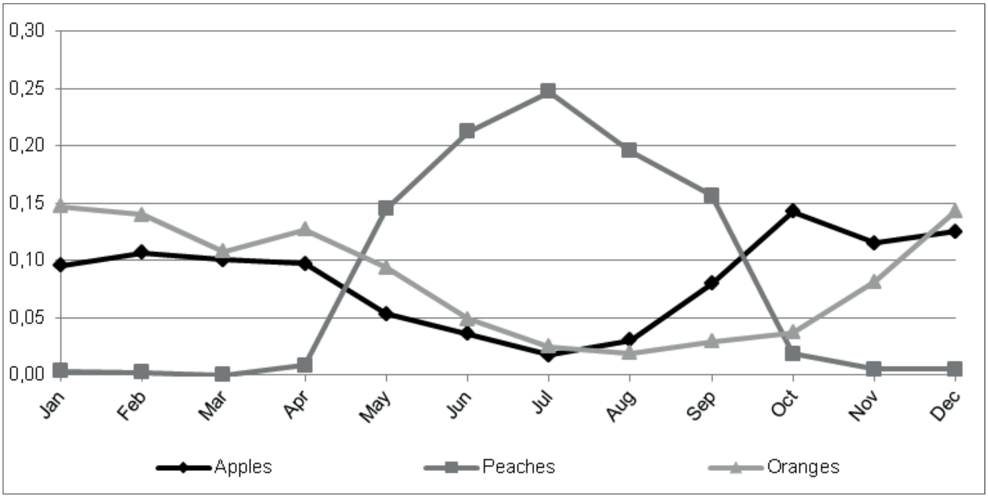
Seeking simplicity and comparability between periods the set of data analysed is reduced to the EAN codes registered in at least one of the relevant weeks in a month. It is implicitly assumed that the sold items only outside the 36 relevant weeks cannot represent valuable products and then can be therefore disregarded. Moreover, the amount of turnover is computed considering only the relevant weeks.

To get a sort of benchmark firstly, the case of well-known categories of seasonal product are investigated. Figure 9 shows the relative amount of turnover (monthly divided by the yearly amount) for Apples, Peaches and Oranges in the 12 months of 2013.

As expected the amount of Peaches turnover is concentrated between May and September while Apples and Oranges are mainly sold in the Fall and Winter period.

It is worth notice that, in order to describe the cycle of life of an EAN code two different aspects have to be considered: (i) if the corresponding product is available for price data collection in a month; (ii) how the amount of turnover is distributed over the whole period. If only the first aspect is considered a product may be classified as continuous because of its availability in each month but the amount of turnover is fully produced in one or two months that induces to classify it as temporary instead.

Figure 9 - Relative amount of turnover for apples, peaches and oranges per 12 months (Torino, 2013)



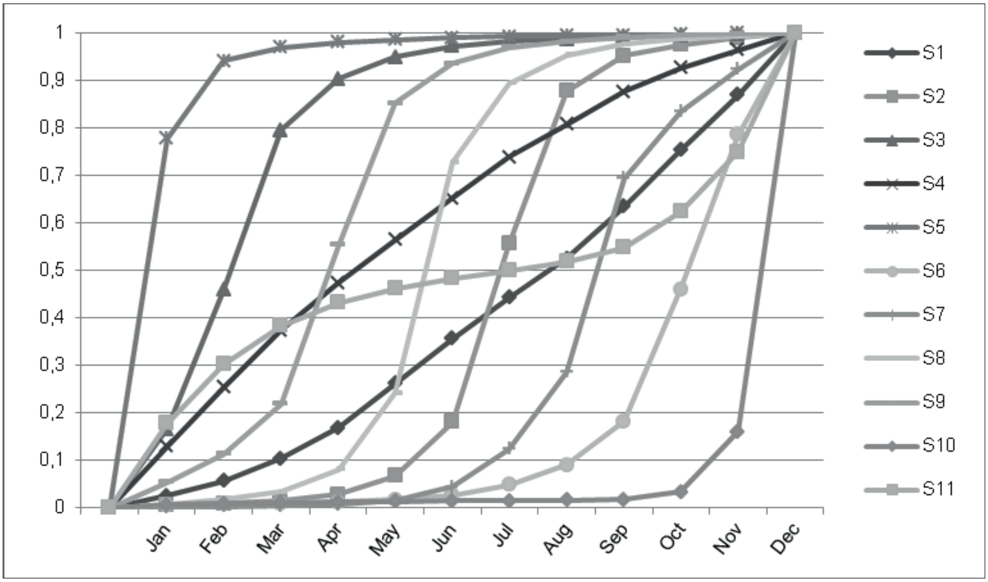
In order to highlight the "standard" different pattern of temporary products, the cumulative over a twelve months period (year 2013) of the relative amount of turnover has been computed. Each elementary item is then characterised by a pattern of 13 increasing values (the first and the last being 0 and 1 for each code). Patterns have been grouped into 11 different clusters according to a *k-mean* model-based procedure[3].

Results are synthetically shown in Figure 10 where the midpoints of each cluster is connected by a line. The purpose of the analysis is to identify standard patterns representing a recognizable cycle of life. The 2 clusters represented by the first line on the left (S5, grey stars) and from the last on the right side (S10, grey rotated squares) can describe the cycle of life of "Christmas products". The line S2 (grey big squares) growing from quite 0 to close to 1 between June and October could describe the pattern of seasonal summer products (as Peaches) while the line S11 (grey squares) growing in the first months and the last ones can be adopted to describe winter seasonal products (Oranges).

3 Software SAS, Fastclus procedure implementing the method called *nearest centroid sorting* introduced in Anderberg, M.R. (1973).

The two lines S1 and S4, growing quite regularly through the year, seem to represent continuous products but, as they have been removed from the data set, may describe "quite continuous" products that are not available in a very few number of months. An EAN code can be therefore classified as seasonal if it presents the same or similar pattern both in 2013 and in 2014.

Figure 10 - Patterns of the cumulative amount of turnover (11 cluster's means) for temporary products (Torino, 2013)



5.5 Seasonality and discontinuity of products

Seasonal products are products sold just at certain times of the year, following a seasonal pattern. According to this definition, it is clear that seasonal products must be detected in the set of *non-continuous* EAN codes, which are sold for less than 12 months during the year.

For the Torino province in the year 2014, this sub-set covers on average only 11.8% of the overall annual turnover, even if it represents 46.5% of series (EAN+outlet code); the peaks of 17.2% and 18.3% of monthly turnover, respectively in April and December, are the first evidence of a seasonal trend in consumptions (Table 7).

Table 7 - Coverage of total turnover per month and total number of EAN codes by item availability (Torino, 2014)

Availability	%Coverage													
	Total turnover per month													Total number of EAN codes (year)
	1	2	3	4	5	6	7	8	9	10	11	12	all	
Less than 12 months	9.3	9.2	9.5	17.2	10.6	11.1	10.8	11.3	11.2	11.0	12.3	18.3	12.0	46.5
12 months	90.7	90.8	90.5	82.8	89.4	88.9	89.2	88.7	88.8	89.0	87.7	81.7	88.0	53.5

Seasonality in consumer prices is caused by a variety of influences, some on the demand side and some on the supply side.

On the demand side, people have different needs depending on the period of the year and on climate conditions. Seasonal consumption patterns arising from this factor normally display about the same behaviour year after year, although the price effects will be modified by supply factors, including possible substitutions. Christmas, Easter, vacation periods and other practices influence the rise and fall in consumption.

On the supply side, the greatest seasonal price changes result from variations in agricultural production, especially among the perishable foods. Consumer demand for these elementary items appears to be quite stable throughout the year, with the result that limited supplies in the seasons when they are not available to determine more elevated prices.

Treatment of seasonal items in CPI is a quite difficult task; their identification requires the availability of at least two yearly collections of weekly/monthly observations in terms of quantities/expenditures, in order to verify whether a periodicity exists.

Analysing data by a group of products is then easy to recognise seasonal movements in the supply and prices of specific *markets* (the lowest level of ECR classification of products, allowing a deeper detail than segments) within consumption segments, although the peaks and the magnitude of seasonal fluctuations may vary widely from year to year: fresh fruits and vegetables, for example, may virtually disappear from the market during certain periods each year. The same problem exists also in some seasonal articles of clothing and in those products which are typically put on the market just for festivities, such as traditional foods or special gift packages.

The identification of seasonal market has been performed on the 2013-2014 collections of monthly total turnover for all products in the Torino province. On the subset of elementary items sold in both years in less than 12 months, the two distinct distributions of the number of contiguous months of sale have been compared in order to isolate, in each segment, the EAN code characterised by a recurring presence approximately in the corresponding period of the considered couple of years.

For those, the annual product turnover in percentage on the total turnover of its market has been calculated and for each market, elementary items have been flagged if they are characterised by a percentage value higher than a determined quantile in the distribution of percentage turnover for all the products in the market.

Finally, the concentration of most flagged EAN codes in a particular market, jointly with a high relative weight of that market on the consumption segment in terms of turnover, identifies it as a "seasonal market".

The described procedure proved to be efficient in the detection of the seasonal component of the annual consumptions curve for the analysed consumption segments. For the group of "Pastry cook products", for example, it has correctly identified the items belonging to the already mentioned market of "Traditional festivity foods" which are sold mainly in Christmas and Easter: while the overall monthly turnover curve depicts a general fall of the consumption from May to August in 2014 (Figure 11), the curve related only to the items in the segment flagged as seasonal clearly shows the opposite trend with a maximum in April and December (Figure 12).

Figure 11 - Percentage of monthly turnover on the total COICOP turnover, COICOP= "Pastry cook products", by all items (Torino, 2014)

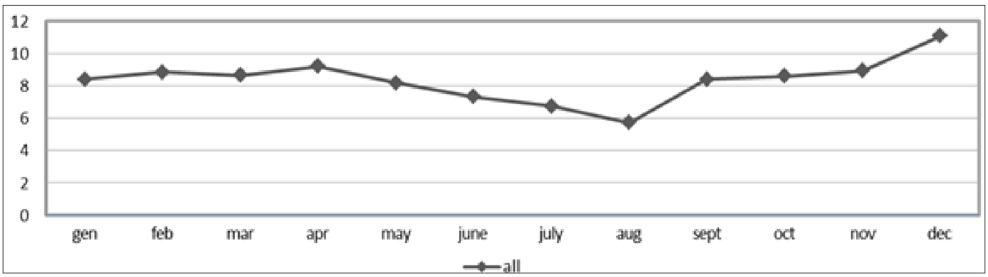
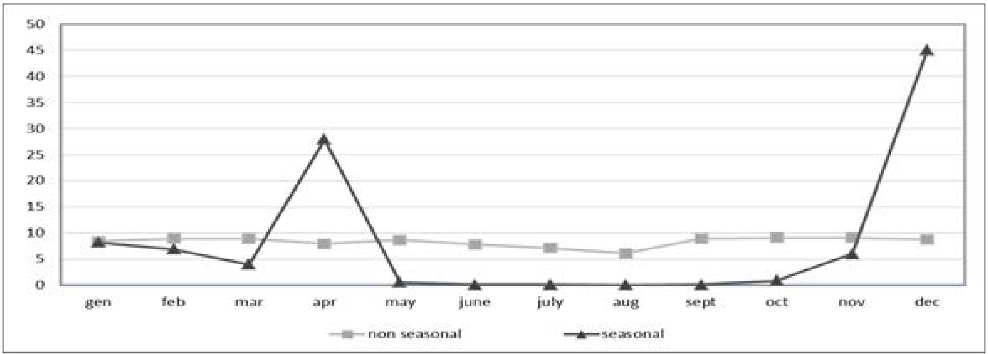


Figure 12 - Percentage of monthly turnover on the total COICOP turnover, COICOP= "Pastry cook products", by seasonality and non- seasonality items (Torino, 2014)



The monthly evolution of expenditures for the complementary set of non-continuous and non-seasonal products is outlined by the grey squares line in Figure 12: here are represented all the items in the consumption segment with a negligible weight with respect to the market or characterised by a certain degree of volatility. Specific assumptions must be made in order to detect the possible sources of volatility, which appears through the fluctuations of sales or by random "entry-exit" patterns during the year.

By applying this procedure, it has been possible to evaluate the coverage of seasonal items for the whole set of product groups: for 2014 in Torino province it consists in 27.7% of the yearly total turnover for the items available less than 12 months, a percentage which grows to 59% in April (Table 8). In the set of non-continuous items, those one flagged as 'not seasonal' count up almost to 87% of the total number of non-continuous series, providing 72.3% of the correspondent turnover.

Table 8 - Coverage of total turnover and total number of EAN codes of seasonal items, by month and item availability (Torino, 2014)

Availability	Seasonality	%Coverage													
		Total turnover per month													Total number of EAN code (year)
		1	2	3	4	5	6	7	8	9	10	11	12	all	
Less than 12 months	yes	24.4	27.9	27.0	59.0	24.0	25.6	16.8	16.5	11.1	13.0	17.0	37.1	27.7	13.1
	no	75.6	72.1	73.0	41.0	76.0	74.4	83.2	83.5	88.9	87.0	83.0	62.9	72.3	86.9
12 months	yes	2.6	3.0	3.0	10.5	2.9	3.1	2.1	2.1	1.5	1.8	2.4	7.2	3.6	2.6
	no	97.4	97.0	97.0	89.5	97.1	96.9	97.9	97.9	98.5	98.2	97.6	92.8	96.4	97.4

While seasonality is fairly known for specific product groups such as the ones already mentioned, seasonal movements are less predictable for particular product sectors: depending on their market share, a periodicity could be not so evident at a first glance just examining the sales trend of the whole segment.

Table 9 figures out the first three COICOP segments, ordered by decreasing percentage, which totalise at least 40% of coverage of the total number of EAN code, respectively for seasonal and non-seasonal items.

The case of low evidence of seasonality is represented by ‘Body lotions’ in the “Personal care” group; with a 14.4% they are the second most sold seasonal items by number of EAN code covered, even if they totalise just 4.7% of the turnover of its COICOP segment. On the contrary, even for those item groups which have a strong seasonality nature (*i.e.* fresh food or clothing), a component of volatility can always be present due to external events.

Table 9 - Coverage of the most sold items on total turnover and total number of EAN codes for discontinuous items, by seasonality and COICOP segment (Torino, 2014)

Seasonality	COICOP segment	COICOP code	% Coverage	
			Total number of EAN codes	Total turnover
Yes	Pastrey cook products	01.1.1.4.2	15.1	15.9
	Body, hand and hair lotions	12.1.3.3.3	14.4	4.7
	Chocolate	01.1.8.3.0	13.8	26.0
	Total		43.3	46.6
No	Articles for personal care, parfumes, make-up	12.1.3.2.2; 12.1.3.3.1; 12.1.3.3.2	21.1	7.2
	Cleaning and maintenance products	05.6.1.1.1; 05.6.1.1.2	8.3	13.1
	Cakes, tarts; ready-made meals; processed vegetables	01.1.7.3.2; 01.1.9.4.0; 01.1.1.4.2	8.1	10.3
	Total		40.1	35.2

6. Experimental framework for sampling from Scanner Data

6.1 Objective and theoretical context

The experiments carried out by the research group were developed in two phases, assuming in the first one only a static population approach (fixed basket) while, in the second, also a dynamic population approach (flexible basket).

The general aim was evaluating the use of SD for the compilation of the elementary price indices (first level of price index calculus on which the subsequent aggregations are based, in the Italian case the consumption segments) from a sampling perspective.

The elementary price indices are computed considering both closed (fixed basket) and open (flexible basket) population: direct indices are built on a fixed basket of products defined at reference time, ignoring new products; direct chain indices are built on a flexible basket that includes all products that disappear or appear (new products) in a year.

In a dynamic context, the bilateral indices are generally calculated on matched-items: only price relatives of items that are sold in two consecutive months enter in the index formulas (flexible basket) (Ivancic *et al.*, 2011). The comparison of two periods, 0 and t, is based on the chain approach. Chain indices take into account the movements of prices within the considered time interval, thus renewing the basket at each sub-interval and, consequently, solve the base change through the change of weights. So, using chain indices the shrinkage effect over time due to the attrition of a fixed basket of products is solved. On the other hand, in the static population, the loss of representativeness of the basket is addressed through the yearly base change of the index and the renewal of the basket. In this context, the comparison of two periods, 0 and t, or binary temporal index, is based on the direct (traditional) approach.

In the dynamic universe, however, period-on-period chaining of weighted indices introduces chain drift, also for superlative Fisher and Törnqvist indices (de Haan *et al.*, 2016). This source of bias, which increases as the time series grows, is due to the non-transitivity of the weighted price indices.

The transitivity of indices is not important in the static universe, as chaining is not required for direct (bilateral) indices, but is more crucial in the dynamic approach.

6.2 Experimental phases

The goal of the first experimental phase was to evaluate the performance of different sample selection schemes of series (references individuated by EAN and outlet codes) and the use of estimators of weighted and unweighted indices for CPI in a static situation. Following a fixed basket method, different samples of series are selected at the beginning of the reference period and followed during the year. In this phase a simplification was used: the implications of the life-cycle of series, seasonality issues and missing data were not taken into account and only panel series were considered as universe for sampling and price index evaluation. The definition of panel data is based on the permanent series concept, which refers to those series with positive turnover for at least one relevant week (the first three full weeks) in each month of the considered year, starting from the December of the previous year.

The population parameters taken in account are three classic aggregation formulas of monthly bilateral price index: Jevons (unweighted), Fisher (ideal) and Lowe (weights from quantities of previous year). In the static population approach, the use of Fisher (superlative) price index formula is undoubtedly the best way to measure price change: Fisher price index is calculated as the geometric mean of the Laspeyres price index and the Paasche price index. Jevons index is an unweighted CPI that uses price information only (it assumes that expenditure shares remain constant), while Fisher and Lowe use also quantity information. These last indices consider expenditure shares at different times (current and reference period) as weights (Gábor and Vermeulen, 2014). Fisher ideal index is thus preferred by economic theory, it uses quantities at different times and allows for substitution effects. The lack of weighting in the Jevons index is a potential source of bias and the opportunity of weighting items "according to economic importance" is supported by the theory of index numbers (de Haan *et al.*, 2016). In a probabilistic sampling context, it has to be specified that the properties of the estimators must also be considered with the properties of the corresponding indices.

In the second phase of the study, some experiments were carried out to highlight the differences between a static and a dynamic population approach in the construction of the elementary price indices. The goal of the experiments was to analyse how some sources of bias can affect the estimates of different index aggregation formulas in both approaches. In the fixed basket approach, bias can be introduced by the reduction in the size of the sample because of disappeared products (shrinkage), by ignoring new products and temporary missing products. In the dynamic population approach, some sources of bias can be related to the matched model and the type of index aggregation formulas utilised. The matched-model based on the exact matching of items sold in two consecutive months does not explicitly account for unmatched new and disappearing items and does not include temporary missing items. Constructing a time series by chaining period-on-period matched-model Jevons indices can avoid chain drift that affects weighted indices. Chain drift occurs if a chained index "does not return to unity when prices in the current period return to their levels in the base period" (Nygaard, 2010; ILO *et al.*, 2004, p. 445). Furthermore, the lack of weighting, the absence of adjustment for quality change and the lack of imputation of temporary missing items are potential sources of index bias (de Haan *et al.*, 2016). The two approaches refer to different sampling schemes: under the static approach, the series is drawn through a two-stage sampling design (outlets and products), while under the dynamic approach, only the selection of outlets is considered.

As in the first phase, the experiments were conducted starting from a panel series, but in this case, artificial populations were generated with products appearing and disappearing (momentarily and permanently). The population parameters here considered are monthly chained bilateral unweighted (Jevons) and weighted superlative indices (Fisher and Törnqvist[4]).

In both experiments, comparison between alternative selection schemes are made for each price index taking the corresponding universe (panel series SD) index value as a benchmark. Indices performance were evaluated in terms of bias for all selection schemes of series. For probability selection schemes, the accuracy (relative bias and sampling variance) of the price indices has been studied in a Monte Carlo simulation scenario. In this context, 500 samples have been selected, according to different sampling designs. Indices

4 Törnqvist index is the weighted geometric average of the price relatives.

variability is measured considering the estimate of the relative sampling error, computed on the estimated indices in the replicated samples. For the sample selection and weighting of price indices, the total annual turnover was taken as reference.

The study was conducted on data relating to the provinces of Torino and Rome. In particular, in the first experimental phase, the probabilistic and non-probabilistic approaches in the selection of series have been investigated assuming as domains of interest three consumption segments (Coffee, Pasta and Mineral Water - COICOP 6 digits) or 88 markets (ECR groups) belonging to the six consumption segments (Coffee, Pasta, Mineral Water, Olive Oil, Spumante and Ice Cream). The SD reference universe is relative to Torino province, 121 outlets and six retail chains (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) available for the year 2014. Scanner data referring to the previous year have been used for the sample selection and weighting of price indices that were based on the total annual turnover of 2013. In the second experimental phase, the study is carried out on Rome province in 2015 and three consumption markets (Short Semolina Pasta, IGP-IGT Italian White Wine, Laundry Bivalent Washing Machine Liquid + Gel).

6.3 Parameters and unbiased estimators

As described above, the parameters of interest taken into account in the experiments are monthly Jevons, Laspeyres, Paasche, Fisher, Lowe and Törnqvist indices.

According to a static population approach, for sake of simplicity, a formalisation of the population parameters considered and the corresponding unbiased estimators are shown in the following scheme (de Haan *et al.* 1999).

Scheme 4 - Price index and sampling estimator

	Population parameter	Sampling estimator
Jevons	$I_J^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{1/n}$	$\hat{I}_J^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{w_i / \sum_{i=1}^{n} w_i}$
Laspeyres	$I_{LA}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) \left(\frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \right)$	$\hat{I}_{LA}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) \left(\frac{p_i^0 q_i^0 w_i}{\sum_{i=1}^{n} p_i^0 q_i^0 w_i} \right)$
Paashe	$I_P^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^0}{p_i^t} \right) \left(\frac{p_i^t q_i^t}{\sum_{i=1}^{n} p_i^t q_i^t} \right)$	$\hat{I}_P^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^0}{p_i^t} \right) \left(\frac{p_i^t q_i^t w_i}{\sum_{i=1}^{n} p_i^t q_i^t w_i} \right)$
Fisher	$I_F^{0,t} = \sqrt{I_{LA}^{0,t} \, I_P^{0,t}}$	$\hat{I}_F^{0,t} = \sqrt{\hat{I}_{LA}^{0,t} \, \hat{I}_P^{0,t}}$
Lowe	$I_{LO}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) \left(\frac{p_i^0 q_i^z}{\sum_{i=1}^{n} p_i^0 q_i^z} \right)$ q_i^z refers to the quantity series in the previous year	$\hat{I}_{LO}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) \left(\frac{p_i^0 q_i^z w_i}{\sum_{i=1}^{n} p_i^0 q_i^z w_i} \right)$
Törnqvist	$I_T^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{1}{2} \left(\frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} + \frac{p_i^t q_i^t}{\sum_{i=1}^{n} p_i^t q_i^t} \right)}$	$\hat{I}_T^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{1}{2} \left(\frac{p_i^0 q_i^0 w_i}{\sum_{i=1}^{n} p_i^0 q_i^0 w_i} + \frac{p_i^t q_i^t w_i}{\sum_{i=1}^{n} p_i^t q_i^t w_i} \right)}$

In the formulas, w_i is the direct weight, *i.e.* the inverse of the inclusion probability of the sampling unit deriving from the sampling design.

Under a dynamic approach, analogous formulas can be expressed for the chain indices, by substituting (0, t) with (t-1, t).

A generic chain index obtained as product of index $I^{0,1}$, $I^{1,2}$,…, $I^{t-1,t}$ referred to sub-intervals s (0,1), (1,2), ...(t-1, t) can be expressed as:

$$I_{chain}^{0,t} = \prod_{s=1}^{t} I^{s-1,s}$$

6.4 Accuracy of price indices

Accuracy of the alternative price indices under different selection schemes of series is evaluated on Monte Carlo simulation scenarios considering the estimates calculated on 500 replicated samples.

Bias and relative sampling error formulas shown below are expressed for a generic parameter (price index) and regarding simulation context.

For a generic estimated index in the *c*-th products group, $\hat{\theta}_c$, the bias can be expressed as

$$B\left(\hat{\theta}_c\right) = E\left[\hat{\theta}_c\right] - \theta_c, \tag{1}$$

in which:

$E\left[\hat{\theta}_c\right]$ is the expected value of the estimated index $\hat{\theta}_c$ in the products group *c*, obtained from 500 samples, and θ_c is the corresponding index value computed on the reference universe (panel series SD).

The relative sampling error of a generic estimated index $\hat{\theta}_c$ in the products group *c* can be expressed by

$$RE\left(\hat{\theta}_c\right) = \frac{\sqrt{Var\left(\hat{\theta}_c\right)}}{\bar{\hat{\theta}}_c}, \tag{2}$$

in which mean and variance of $\hat{\theta}_c$ are calculated on the estimates generated from the selection of 500 samples in the products group *c*.

7. Fixed basket approach: first experimental phase and results

7.1 Probability and non-probability selection schemes

In this experimental phase, probability and nonprobability selection schemes of series in each consumption segment (Coffee, Pasta, Mineral Water) are considered: probability two-stage sampling in the first approach; cut-off and representative elementary item samples in the second approach.

Both for probability and nonprobability sample schemes, series are selected from a sample of outlets (primary stage units - PSU) stratified by chain and outlet type (hypermarket and supermarket). In each stratum, the sample has been allocated proportionally to the turnover. The selection of outlets is carried out in each stratum by simple random sampling (SRS). The outlets' sample size has been fixed at a number of 30 out of 121 outlets of retail trade modern distribution in Torino province. In any case, the series constitute the secondary stage units (SSU).

Nonprobability sampling[5] of series was carried out by selecting series based on cut-off thresholds of covered turnover in the previous year, 2013: two samples are formed with all the series covering respectively the 60 and 80 percent of the total turnover in each of the considered consumption segment in the selected outlets. Moreover, considering the currently used fixed basket approach, a reference selection scheme was defined as selecting the most sold EANs for each representative product in the selected outlets.

For the probability sample, the sample size for SSU is fixed by a sampling rate of 5 percent of the number of EANs in each consumption segment in the sampled outlets. Sample series are selected with probability proportional to size (PPS), in terms of total turnover of the previous year, by adopting Sampford sampling (Sampford, 1967) and Pareto sampling (Rosén, 1997a and 1997b).

Sampford's method is an extension of Brewer's method that selects more than two units from each outlet and without replacement. Units for which

5 The selection of elementary items is made starting from a random sample of outlets. This approach introduce a sampling variance of the estimates determined by the selection of the outlets sample. This component of sampling variability was not taken into account, therefore only the indices bias in non-probability selection schemes is considered.

the initially size measure (turnover during year 2013) is larger than a certain threshold turnover are selected with certainty, where after the inclusion probabilities are calculated for the remainder of the elementary items universe in the sampled outlets. Threshold turnover is defined taking into account the average turnover of outlet, the sampling rate (5%) and the *k* coefficient which can take values greater than 0 and less than 1.

Pareto sampling (PAS) is an order PPS sampling based on the definition of two variables:

- the target inclusion probabilities, $\lambda_{ji} = n_j * s_{ji} \Big/ \sum_{i=0}^{N_j} s_{ji}$,

where n_j is the sample size in the *j*-th outlet (j=1,…,30), determined as product between the sampling rate (5%) and the total number of elementary items in the *j* outlet, N_j, and s_{ji} is the size of *i*-th elementary item ($i=1,...,N_j$) in the *j*-th outlet.

For $\lambda_{ji} > 1$ then let $\lambda_{ji} = 1$.

- the uniform random variable *U* between 0 and 1.

The size measure is associated with each sampling unit and a ranking variable is constructed as a function of these two variables as $Q_{ji} = \frac{U_{ji} * (1 - \lambda_{ji})}{\lambda_{ji} * (1 - U_{ji})}$, in which U_{ji} is a permanent random number (PRN) associated to the *i*-th elementary item in the *j*-th outlet.

In each sampled outlet, elementary units are then sorted in ascending order and the n_j units per outlet with the smallest values of the ranking variable are included in the sample.

In the following scheme the adopted probability and nonprobability selection schemes are synthesised.

Scheme 5 - Probability and nonprobability selection schemes

Selection scheme	Sampling unit	Stratification	Allocation	Selection
One stage - cut-off	Outlet Ean-code	Chain – outlet type	Proportional	SRS Cut-off
One stage – most-sold elementary items	Outlet Ean-code	Chain – outlet type	Proportional	SRS Most-sold elementary items
Two stage sampling	1° Outlet 2° Ean-code	Chain – outlet type	Proportional Fixed sampling rate	SRS Sampford Pareto

7.1.1 Scanner data: operational context

The analyses shown below describe the SD operational framework on which the experiments were carried out: three consumption segments, mineral water (01.2.2.1.0), pasta (01.1.1.6.1) and coffee (01.2.1.1.0) in the Torino province.

Table 10 presents, for each consumption segment, the coverage turnover of relevant week series (B/A) and panel series in relevant weeks (C/B), and the number of panel series.

Table 10 - Total turnover for all series, relevant week series and panel series by consumption segment (Torino, 2014)

Consumption segment	Turnover			%Coverage		Number of panel series
	All series (A)	Relevant weeks all series (B)	Relevant weeks panel series (C)	B/A	C/B	
Coffee	28,622,978	19,665,517	15,692,414	68.7	79.8	9,608
Pasta	26,192,517	17,902,061	13,631,744	68.4	76.2	23,636
Mineral water	26,434,572	18,506,760	16,851,559	70.0	91.1	6,990

Coverage turnover is slightly variable in the consumption segments (minimum 68.4%, maximum 70%) when taking into account the percent ratio (B/A) between the turnover of all series in the relevant weeks and the turnover of all series (52 weeks of the year 2014); coverage turnover is different in the consumption segments when the turnover of the panel series is compared to all series in relevant weeks (C/B) - over 75% for coffee and pasta segments, above 90% for the mineral water segment.

Table 11 describes the turnover covered in the consumption segments in the two cut-off samples with thresholds turnover defined at 60 and 80 percent and the average number of elementary units per outlet considered in the two scenarios.

Table 11 - Percentage and average number of elementary items per outlet covering 60 and 80% of turnover by consumption segment (Torino, 2014)

Consumption segment	Percentage of series		Average number of elementary items per outlet		
	Turnover threshold 60%	Turnover threshold 80%	Total	Covering 60% of turnover	Covering 80% ofturnover
Coffee	16.2	36.1	46	8	17
Pasta	23.4	44.8	114	27	51
Mineral water	12.1	26.3	34	4	9

7.1.2 Main results

In this paragraph some results of the experiments are illustrated: for the mineral water segment the estimates of the twelve-monthly indices achieved through five samples selection are presented; estimates of the monthly Lowe indices are analysed for all consumption segments. The emphasis placed on that index is since it takes full advantage of the information found in the SD.

Tables and figures below present the results of the two Monte Carlo simulations (Sampford and Pareto sampling designs) for monthly Lowe, Fisher and Jevons indices. Besides the outlet selection, a probability or non-probability selection of series has been implemented for comparing the estimates, mainly in terms of bias with respect to the real value of indices.

In seek of brevity, only plots on the mineral water segment have been reported (Figure 13). Figure 13 shows the level estimates of the monthly Lowe, Fisher and Jevons indices computed on probability and non-probability samples and the real value (universe panel series SD) of the corresponding index. Figure 14 shows the level estimates of the monthly Lowe indices computed for coffee and pasta segments.

Irrespective of the consumption segment, the most sold is generally far from the real value and sometimes neither able to catch the trend. The cut-off at 80% is always better than the cut-off at 60% and both are closer to the real value of the index lower is the variability of the segments in terms of prices and turnovers of series.

In general, with the probabilistic selection of elementary items within the selected outlet less biased estimates are obtained. This holds especially under Sampford sampling. Instead, under PAS the results are more biased but more cumbersome because they seem to be dependent on the series variability of the segments in terms of price and/or turnover. The Jevons index represents a kind of exception in this context because the cut-off at 80% and at 60% is less biased than the estimates obtained under Sampford and PAS sampling of series.

Figure 13 - Fisher, Jevons and Lowe indices computed with different selection schemes of series for mineral water segment, year 2014

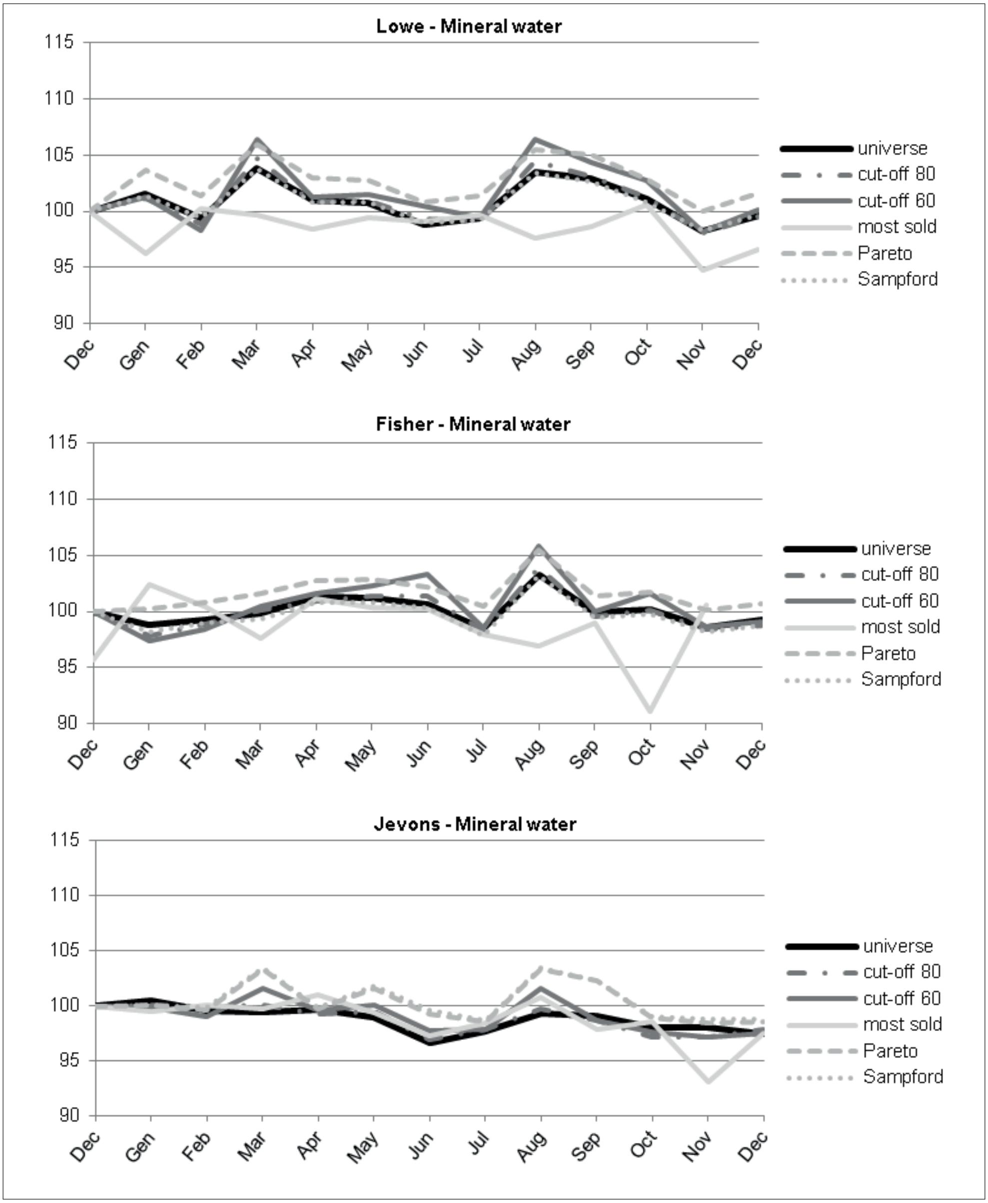
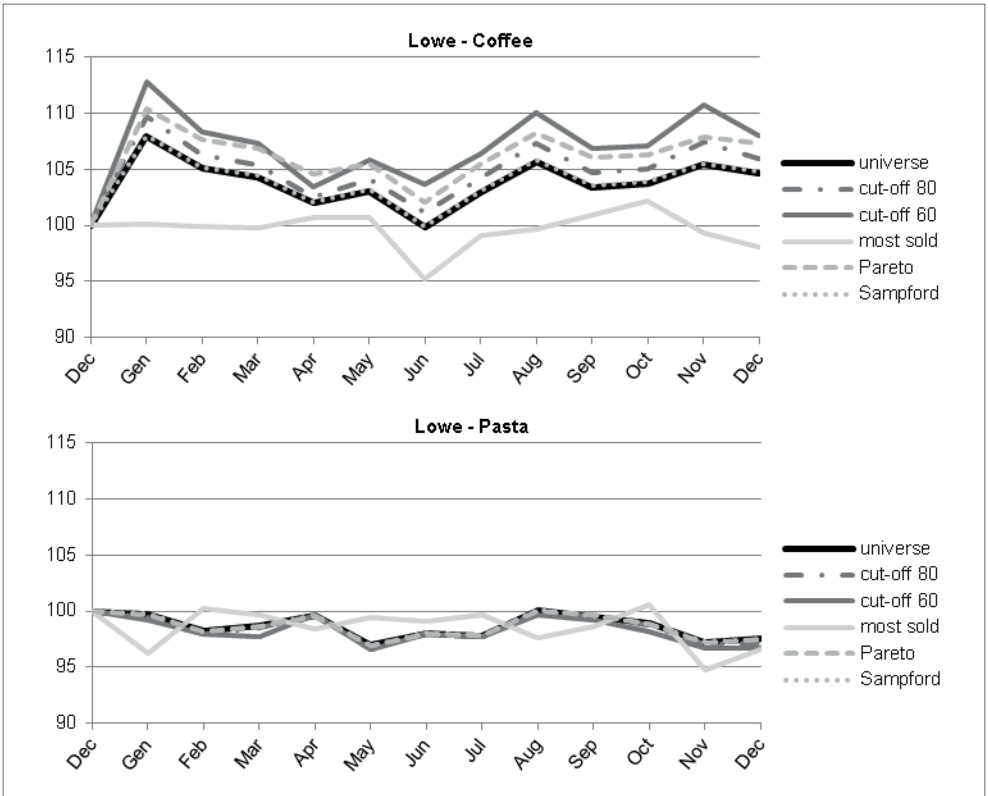


Figure 14 - Lowe indices computed with different selection schemes of series for coffee and pasta segments, year 2014



Looking at Lowe index for the three considered segments, it is possible to see that, under Sampford sampling, Lowe is always unbiased. On the contrary, the most sold item is always far from the real value. Both the cut-offs and also PAS are close or not to the real value depending on the features of the segments, especially concerning the turnover variability of series. In fact, for pasta segment in which the turnover variability is low, the estimates of Lowe index are unbiased irrespective of the method of series selection and apart for the most sold item. In the other two cases the estimates under PAS are in between to those related to the cut-off at 80% and the cut-off at 60%. However, with PAS and with both the cut-off samples the estimates catch the trend of the price index.

In Tables 12 and 13, for each estimated price index, minimum and maximum values assumed by the bias and relative sampling error distributions of the 12 monthly indices are exposed for consumption segment and sampling design.

Table 12 - Bias distribution of monthly Lowe, Fisher and Jevons indices for consumption segment and sampling design

Consumption segment	Sampling design	Love Index		Fisher index		Jevons index	
		Min	Max	Min	Max	Min	Max
Coffee	Sampford	0.05	0.18	-0.40	-0.05	1.99	6.01
	Pareto	2.17	2.60	2.13	2.66	1.95	5.92
Pasta	Sampford	-0.08	0.07	0.09	0.50	-3.12	0.17
	Pareto	-0.11	0.06	-0.34	-0.04	-2.29	0.06
Mineral water	Sampford	-0.26	0.13	-0.71	-0.09	-0.47	4.18
	Pareto	2.74	5.25	1.15	4.81	-0.40	4.77

Table 12 shows that Lowe and Fisher indices present the lowest levels of bias under Sampford sampling in each consumption segment, opposite behaviour can be seen for the Jevons index. Lowe and Fisher indices perform well under Pareto sampling only in the pasta segment.

It is interesting to note that the increase of the bias in PAS design with respect to Sampford, is most conspicuous for Lowe and Fisher indices, mainly in the mineral water segment. The two sampling designs did not show to have a significant impact on the bias of the Jevons index.

Table 13 - Sampling Error distribution of monthly Lowe, Fisher and Jevons indices for consumption segment and sampling design

Consumption segment	Sampling design	Love Index		Fisher index		Jevons index	
		Min	Max	Min	Max	Min	Max
Coffee	Sampford	0.05	0.18	-0.40	-0.05	1.99	6.01
	Pareto	2.17	2.60	2.13	2.66	1.95	5.92
Pasta	Sampford	-0.08	0.07	0.09	0.50	-3.12	0.17
	Pareto	-0.11	0.06	-0.34	-0.04	-2.29	0.06
Mineral water	Sampford	-0.26	0.13	-0.71	-0.09	-0.47	4.18
	Pareto	2.74	5.25	1.15	4.81	-0.40	4.77

The results presented in Table 13 underline very high relative sampling errors for Fischer index and more content for Lowe and Jevons indices in all consumption segments and in both sampling designs.

In the mineral water segment higher relative sampling errors for all indices are found, in particular for Fisher index under Pareto sampling. Considering PAS design is known that the increase of the relative sampling error is most relevant for Lowe index with respect to Jevons index.

Looking both at the bias and relative sampling error distributions of the indices, it follows that, especially in Sampford sampling, Lowe index performs quite well in the coffee and pasta segments but less well in the mineral water segment.

Table 14 shows the coverage probability and width of 95% confidence intervals for Lowe, Fisher and Jevons indices achieved under two sampling designs for each consumption segment.

Table 14 - Coverage probability and width of 95% confidence intervals under Sampford and Pareto sampling with 5% sampling rate for Fisher, Jevons and Lowe indices per consumption segment

Consumption segment	Sampling design	Lowe Index		Fisher Index		Jevons Index	
		Coverage probability %	Width	Coverage probability %	Width	Coverage probability %	Width
Coffee	Sampford	94.70	5.75	95.83	14.58	12.10	6.04
	Pareto	70.17	6.77	92.50	14.40	12.80	9.02
Pasta	Sampford	94.92	4.88	95.38	12.35	68.87	5.41
	Pareto	94.88	4.58	96.00	12.33	76.01	12.20
Mineral Water	Sampford	94.90	10.66	94.35	20.03	78.18	9.29
	Pareto	69.81	11.20	94.68	26.73	77.13	7.60

The confidence intervals (CI) at 95% for Fisher, Jevons and Lowe indices have been derived, through a Monte Carlo method, under Sampford and Pareto with the same sampling fraction (5%).

Crossing the results in Table 14 on the coverage probability and the width of the CIs interesting results are derived. The Fisher index keeps the confidence probability close to the nominal confidence level independently of the consumption segments and the sampling design considered. However, the width of its CIs is usually much wider than those related to the other two indices. Furthermore, Jevons index has narrow CIs among the considered indices due to its low variability, but its coverage probability level is always greatly below the nominal one because it is biased. Finally, Lowe indices has narrow CIs and good coverage probability levels, especially under Sampford sampling. Instead, under Pareto, it has lower coverage probability levels for coffee and mineral water segments. In fact, it seems that Lowe index is affected by bias when computed for segments with high variability of series both in terms of price and turnover.

7.2 Probability sampling designs

Besides comparing probability and non-probability schemes, a deepening was carried out on some probabilistic designs characterised by the use of different criteria of sample allocation, both for outlets and elementary items (EANs), and different selection methods of the sampling units: 1) one-stage stratified sample of EANs; 2) cluster sample of outlets; 3) two-stage sampling with stratification of outlet (PSU) by chain and type (hypermarket and supermarket) and EAN (SSU).

For each sampling design the size of the final sample of EANs was fixed in average at 7,400 to compare the different sampling strategies on equal sizes. The first sampling design was carried out stratifying the EANs by market (ECR group) in each consumption segment. Sample size is allocated among the strata through a Neyman formula, taking into account the variability of price relatives of the EANs in the markets observed in the reference year 2013. Two selection schemes were considered, a simple random sampling (SRS) and with probability proportional to size sampling (PPS) with size equal to turnover (2013). In both selection methods, some sampling units are selected with certainty.

In the second design, cluster sampling, a sample of outlets (14 out of 121 outlets) is selected. Outlets were stratified by chain and type (hypermarket and supermarket). In each stratum, two different allocation of outlets were tested: optimal allocation (Neyman) and proportional allocation based on the turnover of the strata in 2013. Outlets are selected with both SRS and PPS methods. All the EANs in the selected outlets were included in the sample.

Finally, the two-stage sampling design was characterised by a stratification of both PSU and SSU. The stratifications adopted for the PSU and the SSU are the same of the two schemes described above. The PSUs have been allocated through a proportional allocation and selected with PPS based on the turnover of the previous year. The SSU in the selected outlets have been allocated proportionally with the Neyman allocation defined for the stratified sampling and selected with SRS and PPS. To keep in average around 7,400 EAN in this case the sample of outlets has been fixed at a number of 30 out of 121 outlets. In the following scheme the probability sampling designs are synthesised.

Scheme 6 - Probability selection schemes of series

Selection scheme	Sampling unit	Stratification	Allocation	Selection
One stage	EAN-code	Market	Neyman	SRS
				PPS
Cluster	Outlet	Chain – outlet type	Neyman	SRS
				PPS
			Proportional	SRS
				PPS
Two-stage	1° Outlet	Chain – outlet type	Proportional	PPS
	2° Ean-code	Market	Proportional	SRS
				PPS

The comparison among behaviours of the three index aggregation formulas (Jevons, Fisher, and Lowe) has been conducted with the aim of underline the differences among them under the different sampling strategies considered. The same sampling estimator (*i.e.* a plug-in estimator) has been considered for all the index aggregation formulas and the same overall sample size have been drawn under each sampling strategy.

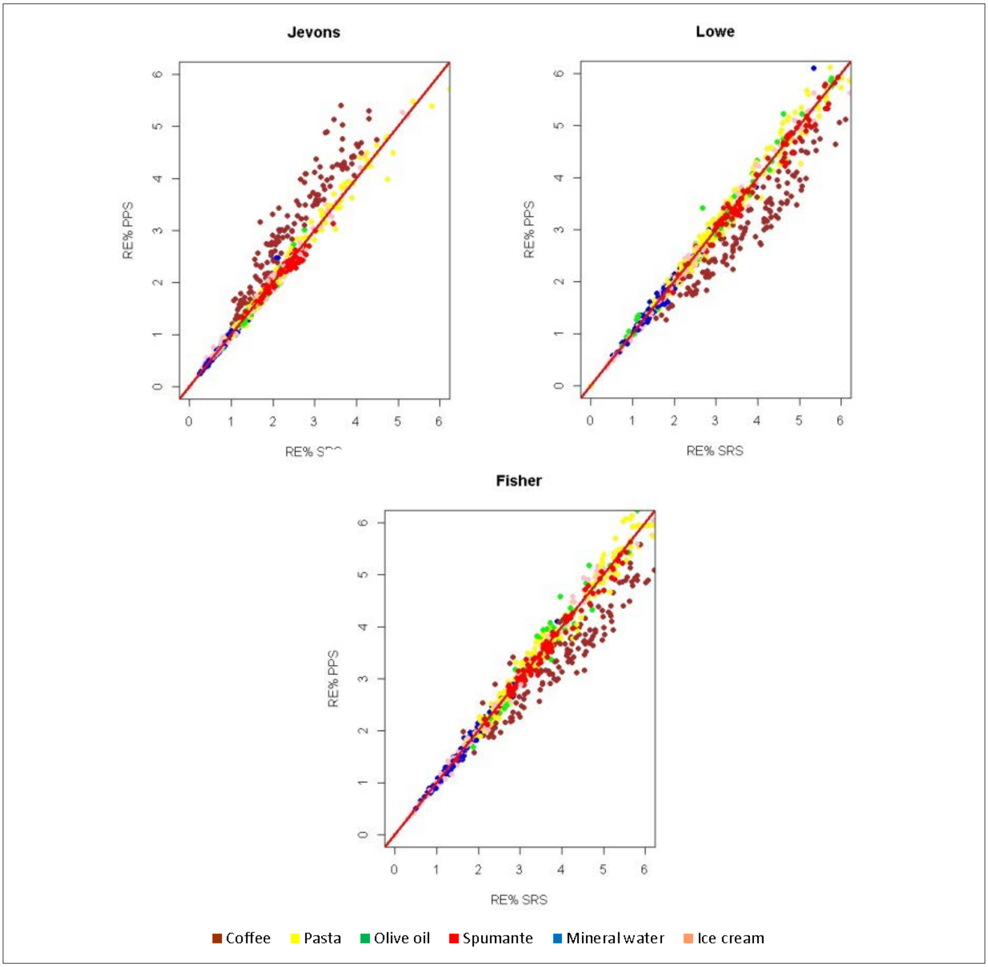
The CPI has been computed at market and consumption segment level for 13 months (from December 2013 to December 2014) for the province of Torino in the 88 markets related to the 6 consumption segments already listed above.

7.2.1 Main results

In the following analyses, relative errors (RE%) of Jevons, Lowe and Fisher estimates are presented under different sampling designs (Scheme 3).

From Figure 15, the estimates, for all the aggregation formulas, are in both cases unbiased (RB –relative bias is approximately equal to 0). With respect to the variability, it possible to see a slight difference among the indices.

Figure 15 - Relative error (RE%) of Jevons, Lowe and Fisher estimates under stratified sampling - SRS and PPS selection methods. Markets in coffee, pasta, olive oil, spumante, mineral water, ice cream. Torino 2014



Looking at Figure 16 the point cloud related to Jevons is usually below those related to Lowe and Fisher, which are almost at the same level.

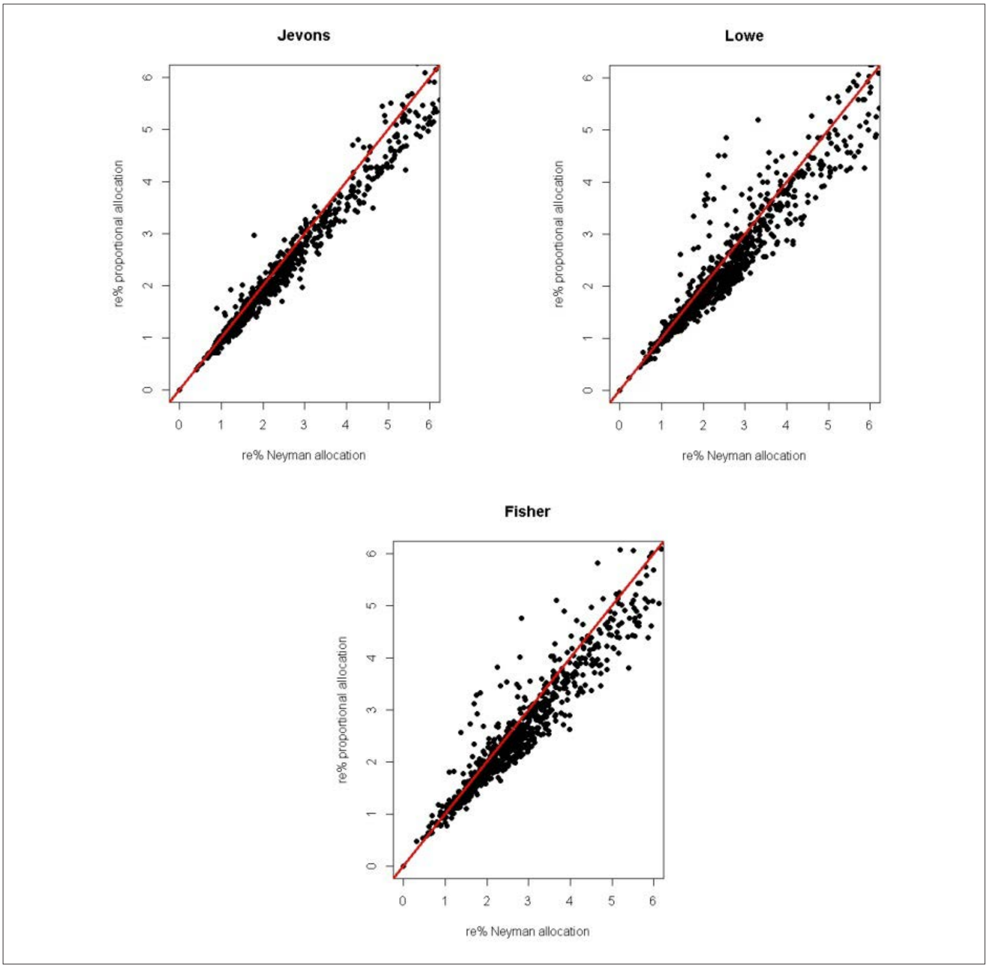
In the figure below, it is possible to notice also a different behaviour of the estimators of the indices with respect to the consumption segments. In particular, for the coffee segment, the estimator of Jevons index is more efficient when EANs are selected under SRS, whilst those for Lowe and Fisher are more efficient when the selection of units is based on PPS. For the other segments, there is no significant differences between the selection methods.

To keep in average around 7,400 EAN the outlets sample has been fixed at 14. In all the scenarios, also, in this case, the estimators of the indices are unbiased. However, when the outlets are selected through a PPS the estimates are more efficient, under both the allocation methods. Instead, between them, the optimal Neyman allocation seems to be less efficient for the outlets. Then the proportional allocation based on turnover of the previous year is preferable.

This advantage is more remarkable when the interest parameters are weighted price indices, such as Lowe and Fisher (Figure 16). In this case, there are no significant differences among consumption segments and among the level of RE% of the estimator of the indices.

As shown in the Figure 16, no significant differences arise using SRS or PPS, probably because most of the variability is at outlet level. This result points the attention to the importance on the allocation and selection method to be used at PSU level and on the size of the PSU sample. In some markets, this could bring an advantage in terms of RE% even if usually two stages sampling implies a higher design effect.

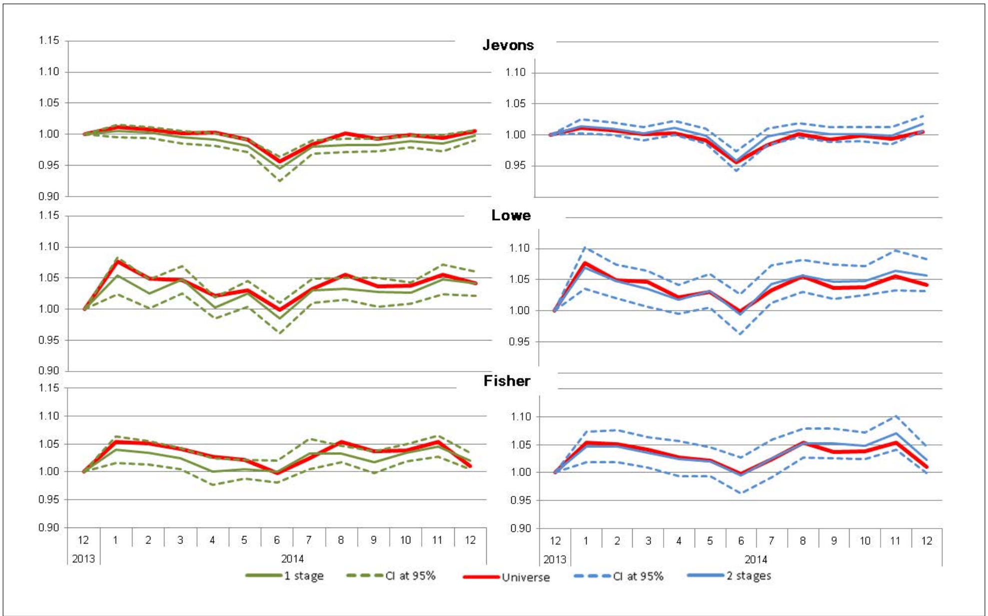
Figure 16 - Relative error (RE%) of Jevons, Lowe and Fisher estimates under stratified one-stage sampling with SRS and PPS selection of outlets. Torino, 2014



Looking at Figure 17 is possible to notice the difference among the indices estimated under the two different sampling strategies. The two sampling strategies compared are stratified one stage, proportional allocation of outlets and PPS selection versus stratified two stages, proportional allocation and PPS selection of outlets and re-proportionated Neyman allocation of EAN in the selected outlets in which EAN have been selected with SRS. The results of estimates on a single sample for the indices in the coffee segment in Torino obtained under the two sampling strategies has compared with the real value (computed on the universe of SD of coffee segment in Torino). All of the estimates seems to catch properly the level and the trend of the related real index. The estimator of Jevons index has in both cases more narrow confidence intervals (CI) with respect to the other two, it means that its RE% is lower. In general, the length of CIs is wider under two stages sampling than under stratified one stage, even if the difference does not seem so large.

In terms of bias with respect to stratified sampling, in one stage and two stages sampling designs, the RB increases slightly, but the estimators can be still considered unbiased.

Figure 17 - Jevons, Lowe and Fisher indices for coffee segment estimated on one sample, confidence interval (CI) of estimates at 95% and real value (computed on the universe of SD). Torino, 2014



8. Fixed and dynamic population: second experimental phase and results

8.1 Objective and description of experiments

The second experimental study aimed to highlight the differences between static and dynamic population approach, using weighted and unweighted indices, and to measure the magnitude of sampling error and bias. The elementary price indices are computed considering both closed and open populations. When assuming a closed population, direct indices are built on a fixed basket of products defined at reference time, ignoring new products (fixed approach), when considering all series of an open population (dynamic approach), direct chain indices are built on matched series of two consecutive months (Ivancic *et al.*, 2011).

The two approaches refer to different sampling schemes. Under the static approach, the series are drawn through a two-stage sampling design, in which the Primary Stage-Units (PSU) are the outlets, the Secondary Stage Units (USS) are the EANs. The outlets, stratified by province, chain and type, are selected with probability proportional to their annual turnover. In each selected outlet, in each market, a sampling fraction of 20% of EANs is selected with probability proportional to annual turnover. Under the dynamic approach, only the selection of outlets is considered.

The experimental study is carried out on Rome province in 2015 and some consumption markets (ECR group). To represent the diversity of the population, three consumption markets different by their features are considered:

- Short Semolina Pasta, low dynamism with respect to products and low variability in prices;
- IGP-IGT Italian White Wine, medium dynamism with respect to products and high variability in prices;
- Laundry Bivalent Washing Machine Liquid + Gel, high dynamism with respect to products and medium variability in prices.

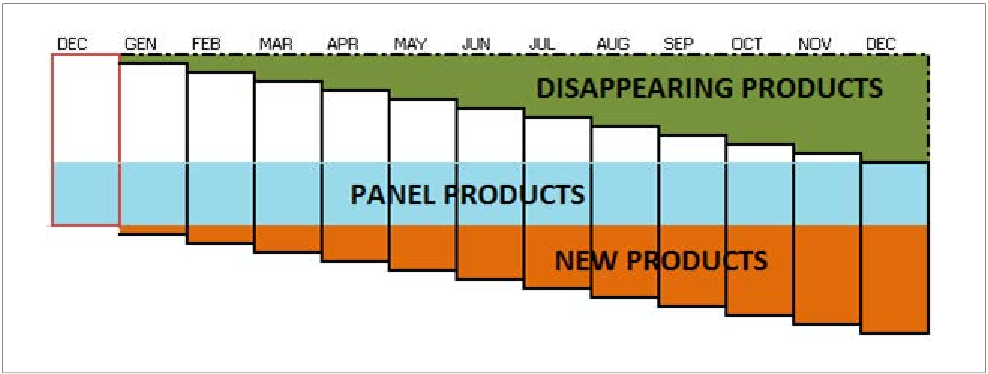
To take into account the variability due to the identification of disappearing, new and temporary missing products and probability sampling, a kind of super-population approach has been implemented.

The panel of products is the reference population. Starting from these data, alternative population have been generated where disappearing and new products, temporary missing, from time to time, are flagged applying survival functions; birth rate and “temporary-missing” rate are estimated on the observed complete data set (from which the panel is extracted as the ‘always present series’). On these populations, the two sampling design is applied and elementary price indices are calculated[6]. The elementary indices considered are Jevons, Törnqvist and Fisher.

Under the dynamic approach, the monthly chained bilateral versions of these indices are used. In this case, the Jevons index is computed on all matched items for two months in a row and on a sub-sample identified by a threshold (matched-model with threshold, “Jevons wT”). The threshold is based on average expenditure shares across two adjacent months; items below the threshold are excluded from the computation. Therefore, an implicit weight is applied (Dutch method) (Van der Grient and de Haan, 2010, 2011).

The following schemes concern the panel series (Scheme 7) and the sub-population considered to decompose the overall survey error deriving from different sources.

Scheme 7 - Panel series, disappearing and new products



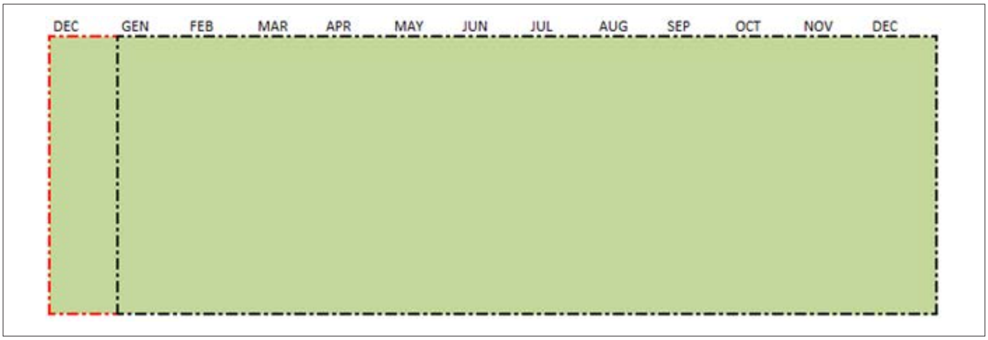
6 500 populations have been drawn and from each population 500 samples have been selected.

In the static approach three sub-populations are considered:

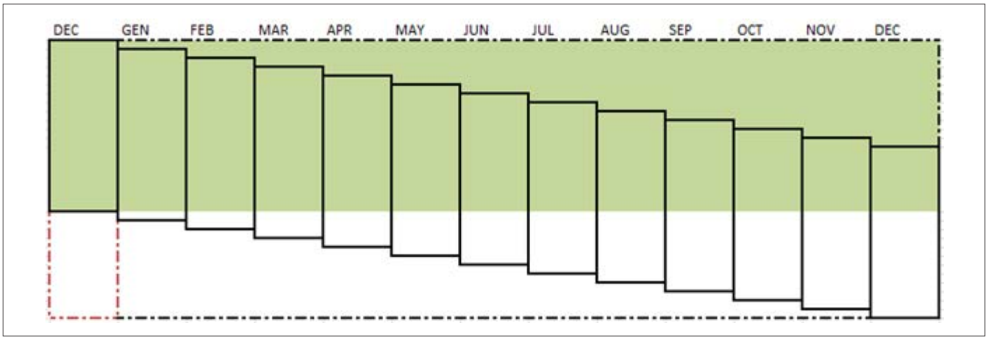
A. Panel series: all the products enter in the computation of indices (Scheme 8);
B. Static population without disappearing products: the products sold in December of the base year are followed during all the year. The new products are not considered by definition. However, disappearing products are assumed to be present (as if replaced) (Scheme 9).
C. "Pure" Static population: the disappearing products are not replaced and the new products are not considered by definition (Scheme 10).

The difference between C and B gives a measure of the error due to the shrinkage, while the difference between B and A quantifies the error due to ignoring the new products. The impact of temporary missing is derived comparing the values of C with and without temporary missing.

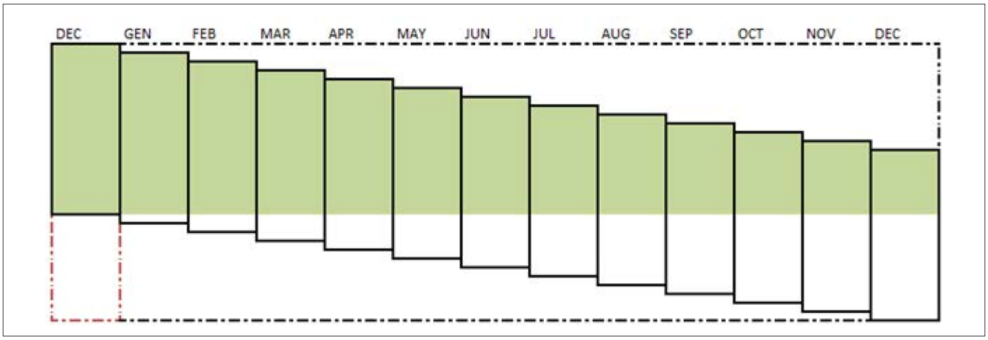
Scheme 8 - Panel series (A)



Scheme 9 - Static population without disappearing products (B)



Scheme 10 - “Pure” static population (C)



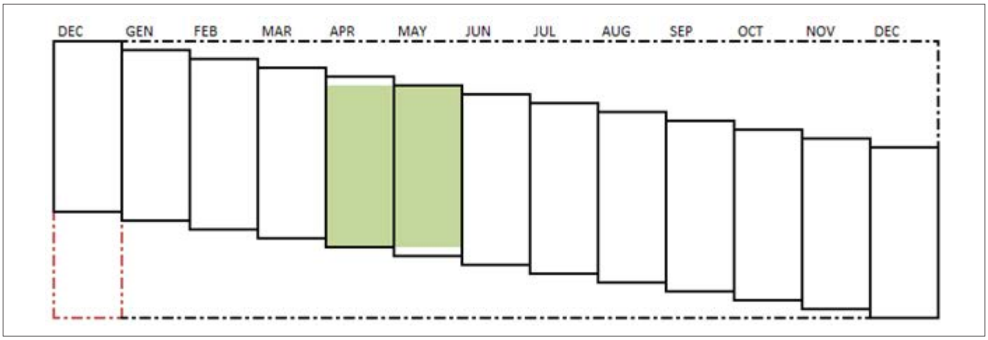
Under the dynamic approach, two sub-populations are considered:

A. Panel series. All the products enter in the computation of indices (Scheme 8).

D. “Pure” Dynamic population: Only the matched products in two months in a row enter in the computation of price indices (Scheme 11).

The difference between D and A can be seen as a difference between a static population and a dynamic one. In this case, the impact of temporary missing is derived by comparing the values of D with and without temporary missing.

Scheme 11 - “Pure” Dynamic population (D)



8.2 Main results

The most meaningful results of this experimental phase are shown to highlight the difference between the static and dynamic approach for the Jevons index.

The following Figure 18 shows the trend of the monthly Jevons index in the segments above listed and in scenarios C and A. The difference between C and A can be seen as a difference between a “pure” static population in which the disappearing products are not replaced and the new products are not considered and a static population (panel series).

Figure 18 shows that in the static approach the bias due to the shrinkage and to ignoring the entering products increases during the year, especially for more dynamic consumption markets.

The effects are small in the “Small semolina Pasta” market and do not affect the variability of estimates, instead, it seems to affect most “IGP-IGT Italian white wine” market that has medium dynamism but high variability in prices.

Figure 18 - Monthly Jevons indices in scenarios A and C for the consumption markets (static approach)

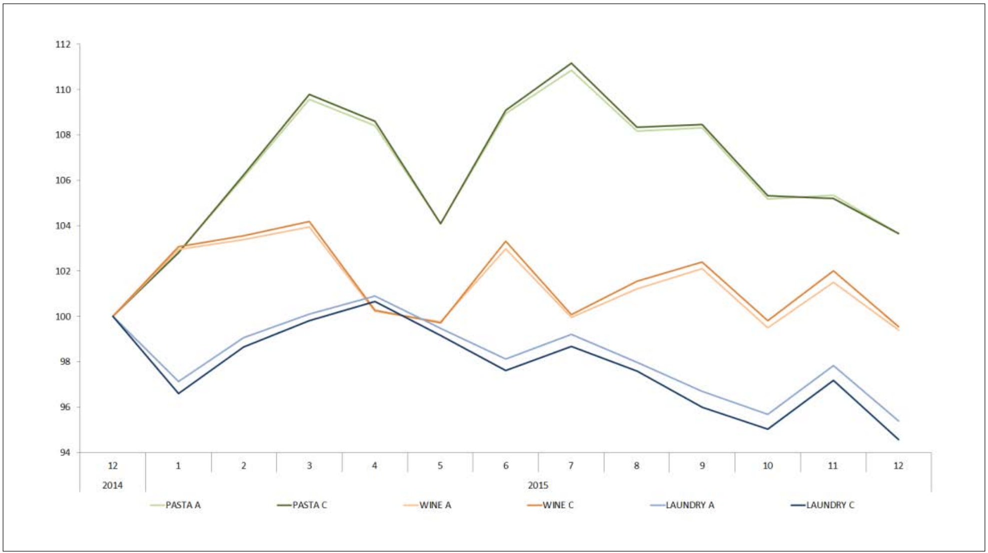
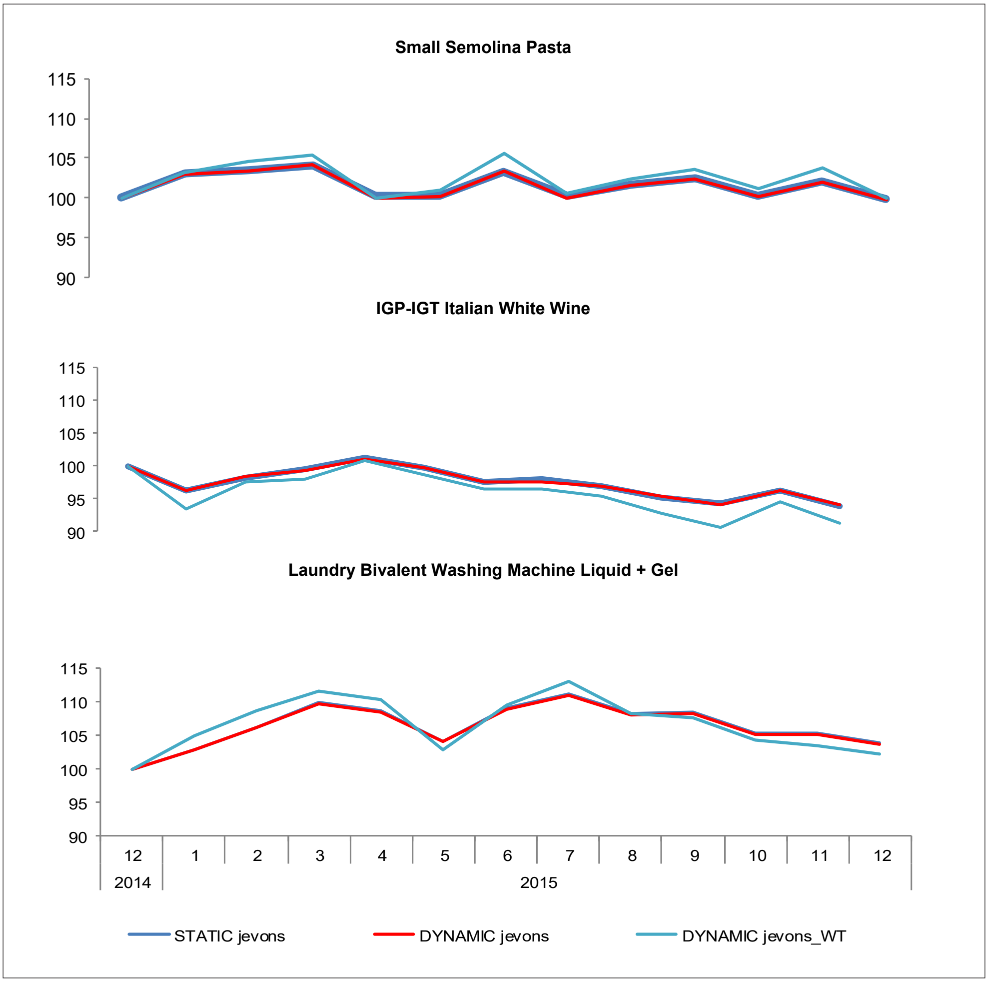


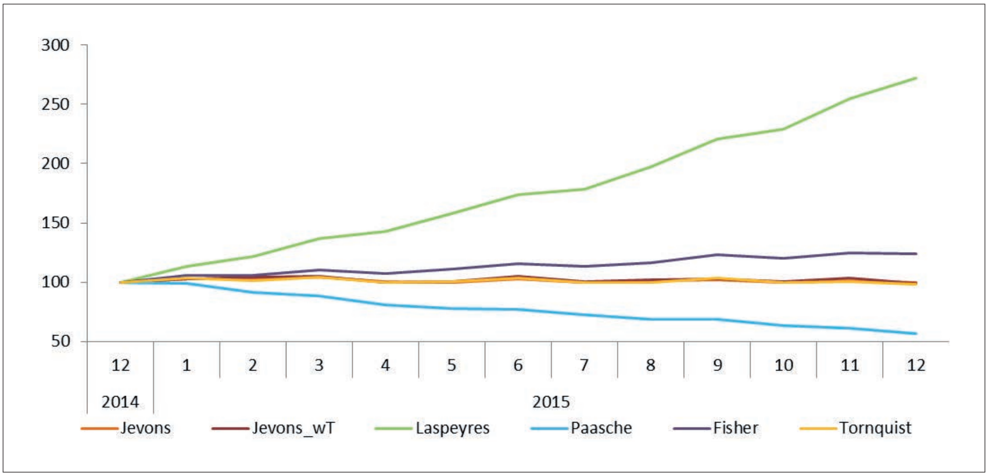
Figure 19 shows a tentative comparison among static and dynamic approach in a close population (scenario A) using the Jevons index with and without threshold. The threshold seems to have a not negligible impact on the levels of the index, mostly for the market with higher variability of prices (IGP-IGT Italian white wine).

Figure 19 - Monthly chained Jevons indices - with and without threshold - in scenarios A under the static and the dynamic approach



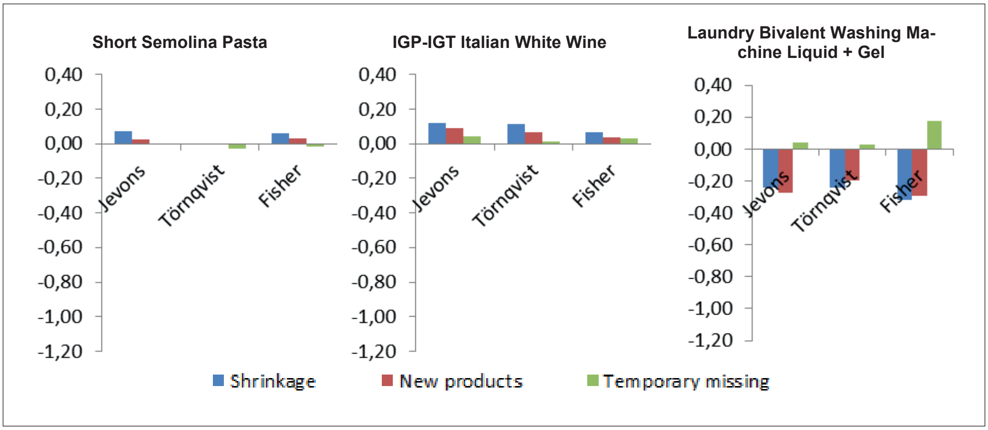
From Figure 20, where weighted and unweighted indices are compared for the “Short semolina pasta” market, it looks clear that the Jevons index does not suffer from chain-drift, while Laspeyres and Paasche index yes. While Jevons index – both with and without threshold – is stable during the year, Laspeyres and Paasche develop along divergent trends. Superlative Fisher and Törnqvist seem to be bounded between Laspeyres and Paasche values, even if Fisher index is not able to tone down the strong increase of the Laspeyres index and it affected by an upward chain-drift. Instead, the Törnqvist index seems to be nearer to Jevons.

Figure 20 - Chained weighted and unweighted price indices under the dynamic approach (Short semolina pasta market)



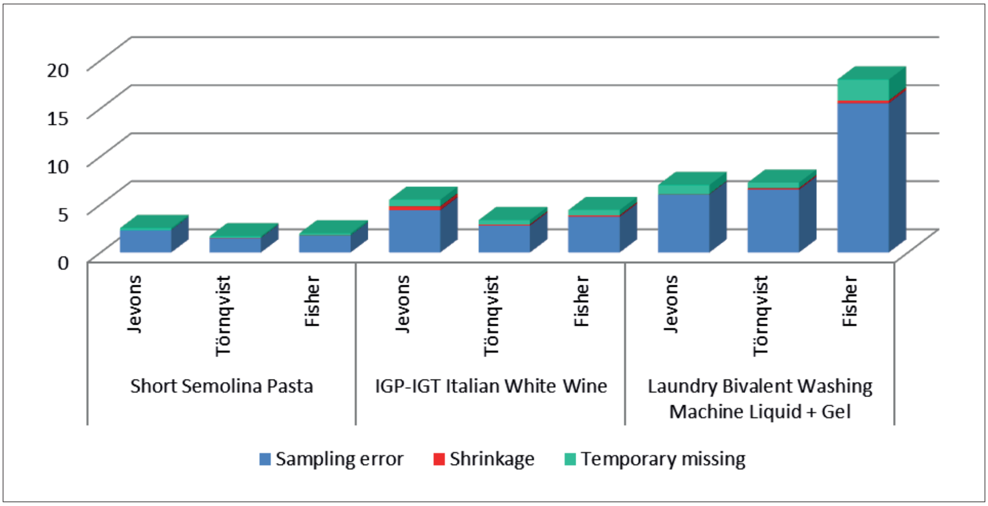
In the following figures, the different sources of bias and sampling variance are analysed in the considered scenarios and approaches.

Figure 21 - Relative percentage differences with respect to the «reference» value of price indices under the fixed approach



In the fixed approach, splitting the bias by sources, we can note in Figure 21 a different behaviour among indices and consumption markets. In general, the bias due to the shrinkage is higher than the bias due to excluding the new products and the presence of temporary missing. Furthermore, it increases with the dynamism of the consumption market.

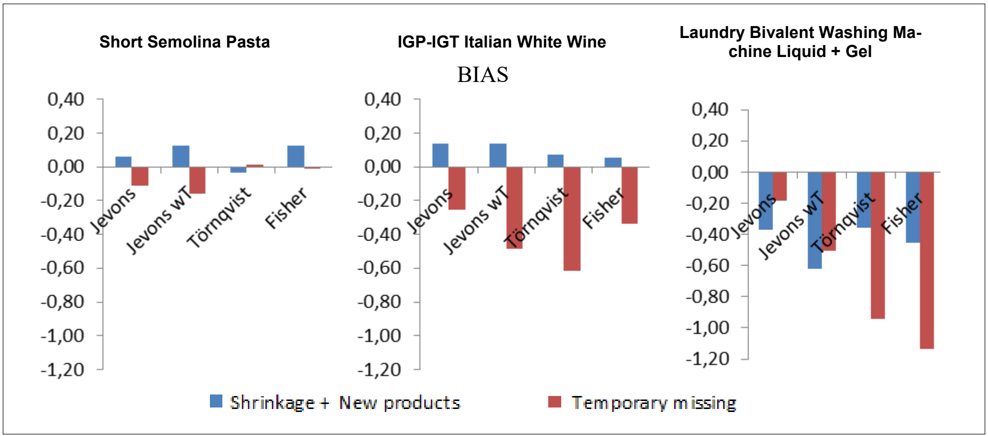
Figure 22 - The increase of percentage relative error due to the shrinkage of products during the year and to the temporary missing under the fixed approach



With respect to the overall variance, as shown in Figure 22, most part is due to the sampling. The shrinkage effect is small in the "Small semolina Pasta" market and does not affect the variability of estimates, instead it seems to affect most "IGP-IGT Italian white wine" market that has medium dynamism but high variability in prices. The bias due to the temporary missing affect more the market with higher dynamism ("Laundry Bivalent Washing Machine Liquid + Gel"). In general, also variance increases with the dynamism of the consumption market.

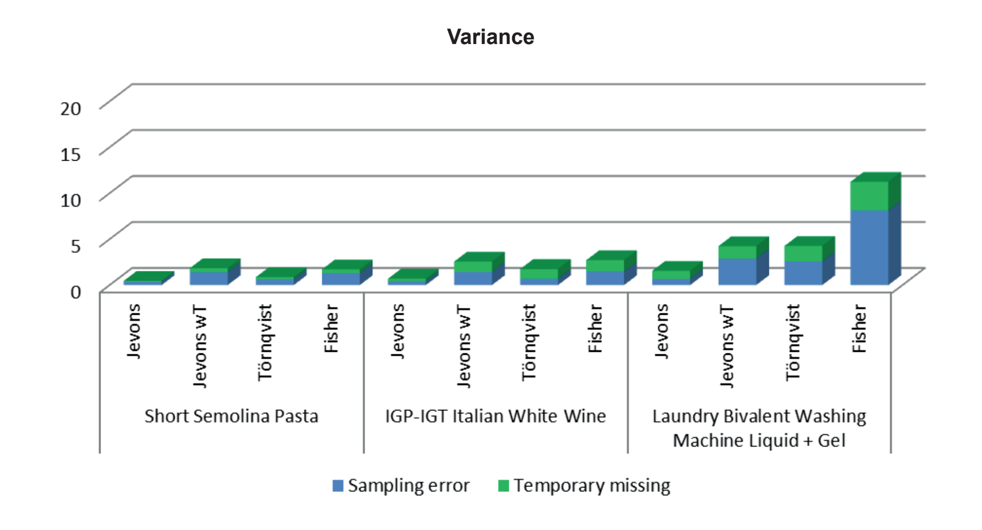
In the dynamic approach, the impact of shrinkage and new product on the bias (Figure 23) is similar to that one under the static approach, while that of temporary missing is higher. This is probably since the size of temporary missing included in the computation of the index under the dynamic approach is much larger than under the static approach.

Figure 23 - Relative percentage differences with respect to the «reference» value of price indices under the dynamic approach



The variance of the indices under the dynamic approach (Figure 24) is very low, due to the sampling design used (cluster instead of two-stage) and the large size of its sample. Also in this case, the variance increase when increasing the dynamism of the market, together with the variance due to temporary missing.

Figure 24 - The increase of percentage relative error due to the shrinkage of products during the year and to the temporary missing under the dynamic approach



In conclusion, also this study highlights the high heterogeneity of the market using scanner data. Sampling and non-sampling error depends on the aggregation formula used and the consumption market features. It seems that the estimators under the dynamic approach are affected by higher bias but much lower variability, especially sampling variability, with those computed under a static approach. This is due to the difference in sample size.

For the same reason, the impact in terms of bias of temporary missing is higher under dynamic approach than under static approach. Further studies can be addressed to derive the sampling and non-sampling error also of multilateral indices which could allow to overcome the chain drift and other issues of weighted indices in the dynamic approach.

9. Sample of outlets from the scanner data of grocery products

9.1 Scanner data in the process of estimation of inflation from 2018

Starting from January 2018 Istat introduces scanner data of grocery products (excluding fresh food) in the production process of estimation of inflation. This innovation concerns 79 indices of an aggregate of products belonging to 5 ECOICOP Divisions (01.02.05.09.12) (Istat, 2019).

In agreement with retail trade chains (RTCs) and with the collaboration of the Association of modern distribution and Nielsen, scanner data for 1,781 outlets (510 hypermarkets and 1,271 supermarkets) of the main 16 RTCs covering the entire national territory are monthly collected by Istat on a weekly basis at the item code level.

For 2018 the compilation of the CPI using scanner data is based on a fixed basket perspective: the use of these data has concerned some channels of the modern distribution, in particular hypermarkets and supermarkets of the main retail chains operating in Italy. Since 2020 Istat has extended the use of scanner data to other channels of the modern distribution (discounts, small sales areas and specialist drug) and has realised the transition to a flexible basket approach.

For the selection of the 2018 sample of outlets (hypermarket, supermarket) a probability design was implemented. Outlets were stratified according to provinces (107), chains (16) and outlet types (hypermarket, supermarket) in 888 strata. Probabilities of selection were assigned to each outlet based on the corresponding turnover value (potential). Concerning the selection of the sample of items, a static approach that mimics the traditional price collection method has been adopted. Specifically, a cut off sample of barcodes (GTINs) has been selected within each outlet/aggregate of products (covering 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover). The products selected in December are kept fixed during the following year. A "tank" of potentially replacing outlets (258) and GTINs (until a coverage of 60% of turnover within each outlet/aggregate) has been detected in order to better manage the possible replacements during 2018.

For 2018, about 1,370,000 price quotes are collected each week to estimate inflation. For each GTIN, prices are calculated taking into account turnover and quantities (weekly price=weekly turnover/weekly quantities). Monthly prices are calculated with the arithmetic mean of weekly prices weighted with quantities. Scanner data (SD) indices of aggregate of products are calculated at outlet level as unweighted Jevons index (geometric mean) of GTINs elementary indices. Provincial SD indices of the aggregate of products are calculated with weighted arithmetic mean of outlet indices using sampling weights. Finally, for each aggregate of products, SD indices and indices referred to other channels of retail trade distribution are aggregated with weighted arithmetic mean using expenditure weights. The sampling design adopted for 2019 followed the same criteria adopted in the previous year for the definition of stratification and allocation but on an updated reference universe. The selection of the new sample was made by maximizing the overlap with the 2018 sample.

In 2020, the extension of the use of scanner data of grocery products to other channels of the modern distribution, including discounts, small sales areas and specialist drug, implied the definition of a new sampling design to take these new outlet types into account in calculating the CPI. In this case the same criteria used for the definition of sample of hypermarkets and supermarkets were adopted except for the stratification. Indeed, outlets were stratified taking into account only provinces and outlet types.

Finally, the overall sample for 2020 was obtained by selecting discounts, small sales areas and specialist drug, maintaining the same selection of hypermarkets and supermarkets used in the previous year. This choice was mainly determined by reasons related to the acquisition of scanner data. The final sample size is 4,073 outlets.

9.2 Sample of outlets from the Nielsen universe

9.2.1 Study of the variability of price indices

To allocate efficiently the sample of outlets (hypermarket, supermarket), the variability of the prices indices in the available 37 provinces and 6 chains has been studied.

Models relating provinces, chains and typology to variability of indices are studied to put more sample in the strata (province * chain * type) with higher variability of the parameter of interest.

The values of monthly price indices with Jevons, Laspeyres and Lowe aggregation formula have been computed for each available outlet of the 37 provinces and the distributions of the standard deviation of outlet monthly indices are analysed. The following figures show the standard deviation of three monthly price indices in different geographical areas.

At the regional level (Figure 25) the standard deviation of the monthly indices is small, except for Lombardia (due to the diversity among Milan and the other provinces) and Veneto (due to the diversity of Venice and the other provinces).

The standard deviation of monthly Jevons indices is more uniform compared to those of Laspeyres and Lowe.

Figure 25 - Standard deviation of monthly price indices by regions, 2015

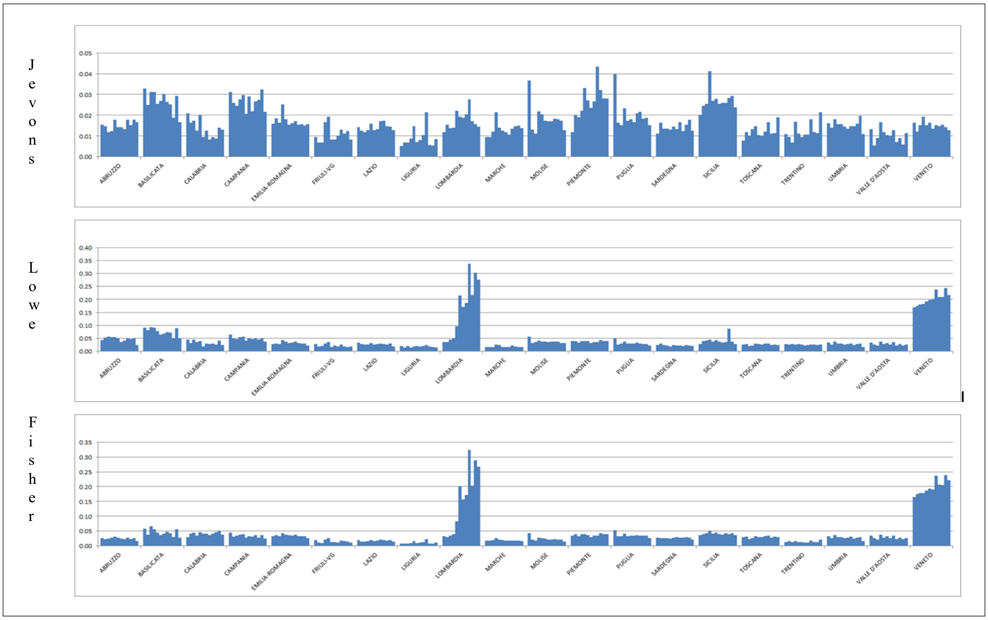


Figure 26 - Standard deviation of monthly indices by provinces, 2015

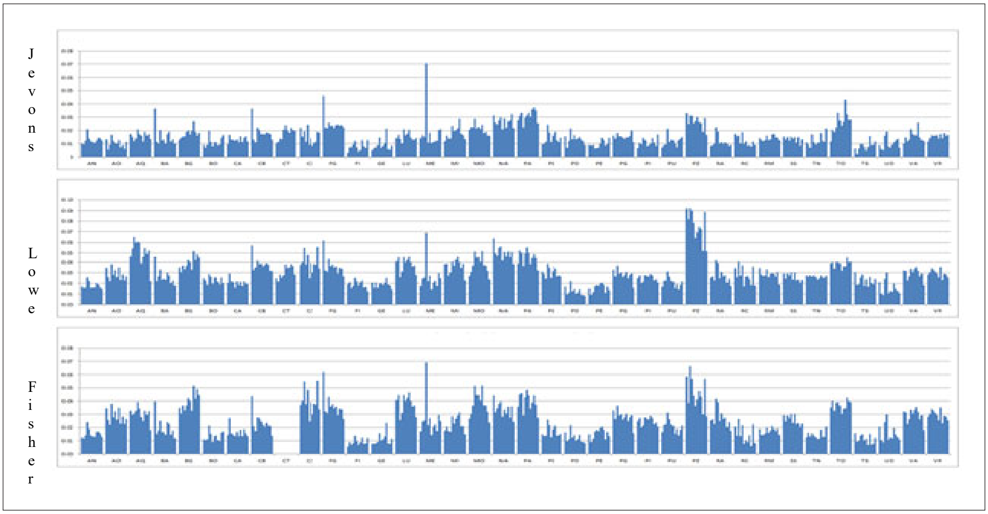
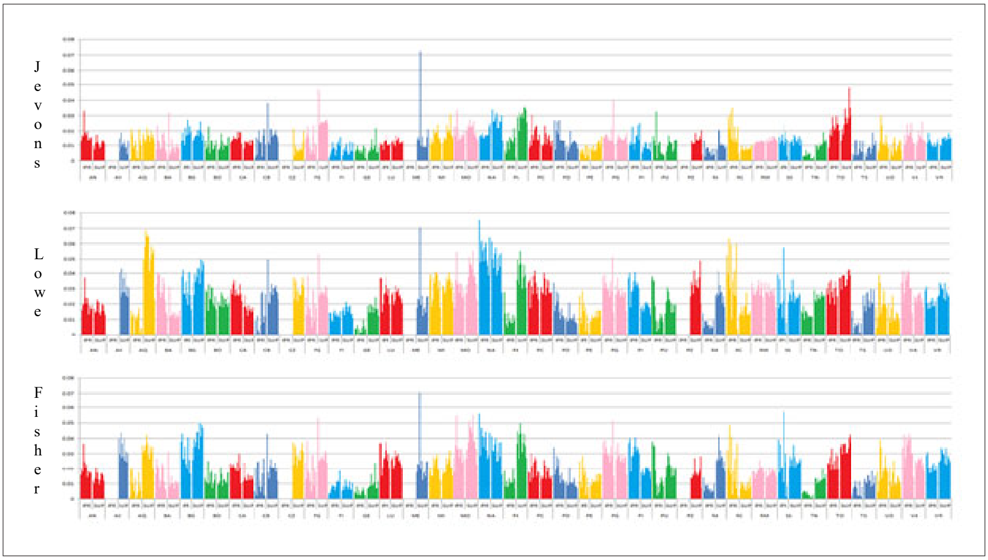


Figure 27 - Standard deviation of monthly indices by provinces and outlet type, 2015



At the provincial level, the standard deviation of the monthly indices increases a little (Figure 26). The three aggregation formulas provide the same results. There is a variability among provinces, but there is not a clear tendency we can use for improving the allocation.

Only in a few cases, the differences, with respect to the variability of monthly indices, between hyper and super-market are clear. But we still cannot assume this relation to address the allocation of outlets (Figure 27).

9.2.2 Sampling design for hypermarkets and supermarkets

The sample of outlets for 2018 was selected by the Nielsen universe represented by 16 chains (out of 25) that gave to Nielsen the authorisation to provide data to Istat, covering nearly 80% of total turnover of hypermarket and supermarket (but not all outlets are available) for in 2017.

Compared to the Nielsen Guide as a whole, considered as a theoretical universe, the sampling frame consists of only the outlets where Nielsen receives the elementary weekly data.

The outlets represent the Primary Stage-Units of the sample design for the selection of references. Outlets are stratified by province, chain and type (hypermarket, supermarket) and the selection is made using PPS, probability proportional to the potential of the outlet.

The objective followed in defining the sample design and the allocation criteria was to obtain a representative sample of the universe of the outlets available in a "neutral" way, *i.e.* not based on assumptions that if not verified could lead to a biased sample. In this way, it is possible to acquire information on the variability of the elementary indices for all the chains and all the provinces, to be used for the re-design of the sample with optimal allocation of the outlets in the strata.

Evaluations have been carried out on the exclusion of strata with very low weight, in terms of potential (turnover), within the province. The following Table 15 shows three hypotheses, with no threshold, with a threshold of 0.5% and 1%, the total number of strata and the number of strata with at least one outlet available.

Table 15 - Distribution of available outlets, strata and strata with available outlets by threshold

Threshold	Available outlet	N. strata	N. strata with available outlets
0	4,733	1,158	913
0.005	4,523	1,094	888
0.01	4,653	1,039	861

Based on the analysis the intermediate option was chosen, the threshold equal to 0.005 (0.5%), which allows at the same time to limit the number of excluded outlets and limit the number of strata with only one outlet.

With a threshold of 0.01, 80 available outlets would be excluded, of which 7 hypermarkets. Their potential is equal to 0.86% of the total, while with a threshold of 0.005, 35 available outlets would be excluded, of which 1 hypermarket (in Lombardy). Their potential is 0.16% of the total.

Allocation between provinces and strata

From the analyses carried out on the data available for the 37 provinces and 6 chains, no clear shreds of evidence emerged on the variability of the price indices that allowed identifying a relationship between variability and potential (the only variable available on all the outlets, including those belonging to the additional chains) from use as an allocation criterion also for the remaining provinces and chains.

For 2018, an allocation based on a compromise between proportionality criteria with respect to potential and with respect to the number of outlets in the strata is defined. The attribution of an excessive weight to the potential would have led to include almost all hypermarkets in the sample, limiting the number of supermarkets.

The allocation of 2,100 outlets, defined ex-ante, was obtained in two steps and is based on the total potential and number of outlets referring to the universe of outlets available and not (theoretical universe).

The number of outlets by the province has been defined on a compromise between three allocations:

- a uniform allocation among the 107 provinces (with a weight of 0.6);
- an allocation proportional to the number of outlets in the provinces (with a weight of 0.3);
- an allocation proportional to the total potential in the provinces (with a weight of 0.1).

The number of outlets by type (hypermarket, supermarket) and chain within each province have been defined based on a compromise between two allocations:

- an allocation proportional to the total number of outlet (with a weight of 0.65);
- an allocation proportional to the total potential of the PV (with a weight of 0.35).

The sample sizes per stratum were therefore defined on the theoretical universe and then adjusted to take into account the number of outlets available. Figure 28 shows the distribution of the sample of outlets in the Italian regions.

Figure 28 - Distribution of the sample of outlets (hypermarkets, supermarkets) by region. Year 2018

For 2019 the same sampling design was used updating only the reference universe (Nielsen 2018).

The selection of the sample of outlets (hypermarket, supermarket) was made by maximizing the overlap of the new sample with the sample of the previous year. The overlap achieved amounts to almost 85%.

The following table shows the distributions of the sample of outlets selected at the regional level.

Table 16 - Distribution of the sample of outlets (hypermarkets, supermarkets) by region. Year 2019

Region	N. outlet
Piemonte	169
Valle D'Aosta/*Vallée d'Aoste*	7
Lombardia	327
Trentino-Alto Adige/*Südtirol*	42
Veneto	177
Friuli-Venezia Giulia	75
Liguria	76
Emilia-Romagna	185
Toscana	168
Umbria	40
Marche	89
Lazio	137
Abruzzo	70
Molise	26
Campania	108
Puglia	105
Basilicata	16
Calabria	76
Sicilia	165
Sardegna	93
Totale	**2,151**

9.2.4 Sampling design for discounts, small sales areas and specialist drug

In 2019, Nielsen's updated the outlet list, which now contained also discounts, small sales areas and specialist drug, as well as hypermarkets and supermarkets, already present in previous years. The sample allocation of outlets, also for the 2020 sample, has been studied according to the potential.

The sample of new outlet types for 2020 was selected by the Nielsen universe (year 2019). Compared to the Nielsen Guide as a whole, considered as a theoretical universe, the sampling frame consists of only the outlets from which Nielsen receives elementary weekly data.

The outlets are the Primary Stage-Units of the sample design for the selection of references. Outlets are stratified by province and type (discounts, small sales areas and specialist drug) and the selection is performed using PPS, probability proportional to the potential of the outlet.

The sample size, defined ex-ante, of 4,000 outlets was obtained (analogously to what happened in the last survey) in two steps and is based on the total potential and number of outlets referring to the universe of outlets available and not (theoretical universe).

The number of sample outlets by province has been defined on the basis of a compromise between three allocations:

- a uniform allocation among the 107 provinces (with a weight of 0.6);
- an allocation proportional to the number of outlets in the provinces (with a weight of 0.2);
- an allocation proportional to the total potential in the provinces (with a weight of 0.2).

The number of outlets by type (discounts, small sales areas and specialist drug) within each province has been defined based on a compromise between two allocations:

- an allocation proportional to the total number of outlet (with a weight of 0.42);
- an allocation proportional to the total potential of the PV (with a weight of 0.58).

In the strata in which there are fewer outlets than those allocated, all units in the stratum have been selected.

The final sample consists of 1,951 outlets belonging to discounts, small sales areas and specialist drug. The following table shows the distributions of the sample of outlets selected at the regional level.

Table 17 - Distribution of the sample of outlets (discounts, small sales areas and specialist drug) by region. Year 2020

Region	N. outlet
Piemonte	136
Valle D’Aosta/*Vallée d'Aoste*	12
Lombardia	220
Trentino-Alto Adige/*Südtirol*	46
Veneto	126
Friuli-Venezia Giulia	60
Liguria	80
Emilia-Romagna	138
Toscana	154
Umbria	36
Marche	80
Lazio	134
Abruzzo	68
Molise	32
Campania	131
Puglia	141
Basilicata	39
Calabria	69
Sicilia	168
Sardegna	81
Total	**1,951**

10. Conclusions

The results of the experimental phases carried out using the first scanner data sets provided to Istat allowed us to individuate the most efficient sampling scheme for the selection of outlets and references in the static approach and to obtain a measure of sampling error and bias in the dynamic approach. The first experiments produced interesting results regarding the performance of sampling schemes and index formulas in a closed population context and fixed approach. They lead to the conclusion that probability sampling is the better choice in this context.

The possibility of switching to a dynamic approach requires, from an economic perspective, to deal with some complex issues. In fact, weighted indices would enable us to exploit better the potential of scanner data for the estimate of an elementary index, but they are affected by different drawbacks, first of all, the chain drift.

The international debate on the use of scanner data and therefore of a dynamic population approach is currently centred on the issue of how to resolve the chain drift problem related to the chaining of weighted price indices (also Fisher and Törnqvist), while at the same time maximizing the number of matches in the data. A solution to the chain-drift issue might be the construction of weighted transitive multilateral price indices, which are free from chain drift by definition, but other issues must be solved to ensure that previously published index numbers will not be revised. Chaining matched-model superlative indices are recommended in the ILO Manual (ILO, 2004) for the satisfactory properties that characterize them.

The outlined second experimental phase will provide evidences on the pros and cons of the two approaches, highlighting in particular empirical and theoretical drawbacks of the dynamic approach which is the one that Istat introduced in 2020. Therefore, the Institute has made a gradual transition from an approach based on a fixed basket to an approach based on a flexible basket. From this perspective, Istat is currently participating in the international debate focussed on the search for solutions that aim to the full use of the information contained in the SD and consequently to the use of weighted elementary price indices for the calculation of the CPI.

References

Anderberg, M.R. 1973. *Cluster Analysis for Applications*. Cambridge, MA, U.S.: Academic Press.

Chessa, A.G., J. Verburg, and L. A. Willenborg. 2017. "Comparison of Price Index Methods for Scanner Data". Paper presented at the *15th Meeting of the Ottawa Group*. Eltville, Germany, 10th - 12th May 2017.

de Haan, J., E. Opperdoes, and C.M. Schut. 1999. "Item selection in the Consumer Price Index: Cut-off versus probability sampling". *Survey Methodology*, 25(1): 31-41.

de Haan, J., and H.A. van der Grient. 2011. "Eliminating chain drift in price indexes based on scanner data". *Journal of Econometrics*, 161: 36-46.

de Haan, J., L. Willemborg, and A.G. Chessa. 2016. "An overview of price index methods for scanner data" (preliminary draft).

Feldmann, B. 2015. "Scanner data: current practice". Presentation at the *Workshop Scanner Data*. Roma, Italy: Italian National Institute of Statistcs – Istat, 1st - 2nd October 2015.

Gábor, E., and P. Vermeulen. 2014. "New evidence on elementary index bias". European Central Bank, *Working paper series*, N. 1754/ December 2014.

International Labour Office - ILO, International Monetary Fund - IMF, Organisation for Economic Co-operation and Development - OECD, Statistical Office of the European Communities - Eurostat, United Nations, and The World Bank. 2004. *Consumer price index manual: Theory and Practice*. Geneva, Switzerland: ILO Publications.

Ivancic, L., W.E. Diewert, and K.J. Fox. 2011. "Time Aggregation and the Construction of Price Indexes". *Journal of Econometrics*, 161(1): 24-35.

Nygaard, R. 2010. "Chain drift in a monthly chained superlative price index". *Workshop on scanner data*. Geneva, Switzerland, 10th May 2010.

Norberg, A. 2014. *Sampling of scanner data products offers in the Swedish CPI. Draft version 8*. Solna and Örebro, Sweden: Statistics Sweden.

Italian National Institute of Statistics - Istat. 2019. "Prezzi al Consumo. Dati defnitivi". *Statistiche flash*. Roma, Italy: Istat. https://www.istat.it/it/archivio/226109.

Rais, S. 2008. "Outlier detection for the consumer price index". In *Proceedings of the Survey Methods Section of the SSC Annual Meeting*. Ottawa, Ontario, Canada, 25th - 27th May 2008.

Rosén, B. 1997a. "Asymptotic theory for order sampling". J*ournal of Statistical Planning and Inference*, 62(2): 135–158.

Rosén, B. 1997b. "On sampling with probability proportional to size". *Journal of Statistical Planning and Inference*, 62(2): 159–191.

Saidi, A., and S. Rubin Bleuer. 2010. "Detection of outliers in the Canadian consumer price index". Paper presented at the *Conference of European Statisticians*. Ottawa, Ontario, Canada, 16th - 18th May 2015.

Sampford, M.R. 1967. "On sampling without replacement with unequal probabilities of selection". *Biometrika*, 54(3-4): 499–513.

Van der Grient, H., and J. de Haan. 2010. "The Use of Supermarket Scanner Data in the Dutch CPI". Paper presented at the *Joint ECE/ILO Workshop on Scanner Data*.

Van der Grient, H., and J. de Haan. 2011. "Scanner Data Price Indexes: The "Dutch" Method versus Rolling Year GEKS". Paper presented at the *12th Meeting of the Ottawa Grou*p. Wellington, New Zealand, 4th - 6th May 2011.

Vermeulen, B.C., and H.M. Herren. 2006. "Rents in Switzerland: sampling and quality adjustment". Paper presented at the *11th Meeting of the Ottawa Group*. Neuchâtel, Switzerland, 27th – 29th May 2006.